Chi-square:

$$\chi^2 = \sum \left[\frac{1}{\sigma_i^2} (y_i - a - bx_i)^2 \right]$$

Least-squares fitting procedure: Minimize χ^2 with respect to each of the coefficients simultaneously.

Coefficients of least-squares fitting:

$$a = \frac{1}{\Delta} \left(\sum \frac{x_i^2}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} \right)$$

$$b = \frac{1}{\Delta} \left(\sum \frac{1}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} \right)$$

$$\Delta = \sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2$$

Estimated variance s^2:

$$\sigma^2 \simeq s^2 = \frac{1}{N-2} \Sigma (y_i - a - bx_i)^2$$

Statistical fluctuations:

$\sigma_i^2 \simeq y_i$ raw data counts

Uncertainties in coefficients:

$$\sigma_a^2 \simeq \frac{1}{\Delta} \sum \frac{x_i^2}{\sigma_i^2} \qquad \sigma_b^2 \simeq \frac{1}{\Delta} \sum \frac{1}{\sigma_i^2}$$

EXERCISES

6-1 Fit the data of Example 6-2 as if all of the data had equal uncertainties $\sigma_i = \sigma$.

6-2 How would you go about solving the simultaneous equations of Equations (6-15)?

6-3 Fit the data of Example 6-1 as if all the uncertainties followed the Poisson distribution $\sigma_i^2 \simeq T_i$.

6-4 Derive Equations (6-25).

6-5 Compare the discrepancies Δ_i of Example 6-1 with the experimental uncertainty s. How much larger than s is the largest value of Δ_i? How probable is such a discrepancy?

6-6 Show that Equations (6-12) reduce to Equations (6-9) if $\sigma_i = \sigma$.

6-7 Derive a formula for making a linear fit to data with an intercept at the origin $y = bx$.

CHAPTER SEVEN

CORRELATION PROBABILITY

7-1 LINEAR-CORRELATION COEFFICIENT

Let us assume that we have made measurements of pairs of quantities x_i and y_i. We know from Chapter 6 how to make a least-squares fit to these data for a linear relationship, and in the next chapters we will consider fitting the data with more complex functions. But we must also stop and ask whether the fitting procedure is justified, whether, indeed, there *exists* a physical relationship between the variables x and y. What we are asking here is whether or not the variations in the observed values of one quantity y are *correlated* with the variations in the measured values of the other quantity x.

For example, if we were to measure the length of a metal rod as a function of temperature, we would find a definite and reproducible correlation between the two quantities. But if we were to measure the length of the rod as a function of time, even though there might be fluctuations in the observations, we would not find any significant reproducible long-term relationship between the two sets of measurements.

On the basis of our discussion in Chapter 6, we can develop a quantitative measure of the degree of linear correlation or the probability that a linear relationship exists between two observed quantities. We can construct a linear-correlation coefficient r which will indicate quantitatively whether or not we are justified in determining even the simplest linear correspondence between the two quantities.

Reciprocity in fitting x vs. y Our data consist of pairs of measurements (x_i,y_i). If we consider the quantity y to be the dependent variable, then we want to know if the data correspond to a straight line of the form

$$y = a + bx \tag{7-1}$$

We have already developed the analytical solution for the coefficient b which represents the slope of the fitted line given in Equation (6-9),

$$b = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N\Sigma x_i^2 - (\Sigma x_i)^2} \tag{7-2}$$

where the weighting factors σ_i have been omitted for clarity. If there is no correlation between the quantities x and y, then there will be no tendency for the values of y to increase or decrease with increasing x, and, therefore, the least-squares fit must yield a horizontal straight line with a slope $b = 0$. But the value of b by itself cannot be a good measure of the degree of correlation since a relationship might exist which included a very small slope.

Since we are discussing the interrelationship between the variables x and y, we can equally well consider x as a function of y

and ask if the data correspond to a straight line of the form

$$x = a' + b'y \tag{7-3}$$

The values of the coefficients a' and b' will be different from the values of the coefficients a and b in Equation (7-1), but they are related if the variables x and y are correlated.

The analytical solution for the inverse slope b' is similar to that for b in Equation (7-2).

$$b' = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N\Sigma y_i^2 - (\Sigma y_i)^2}$$

If there is no correlation between the quantities x and y, then the least-squares fit must yield a horizontal straight line with a slope $b' = 0$ as above for b.

If there is complete correlation between x and y, then there exists a relationship between the coefficients a and b of Equation (7-1) and between a' and b' of Equation (7-3). To see what this relationship is, we rewrite Equation (7-3)

$$y = -\frac{a'}{b'} + \frac{1}{b'}x = a + bx$$

and equate coefficients.

$$\begin{aligned} a &= -\frac{a'}{b'} \\ b &= \frac{1}{b'} \end{aligned} \tag{7-4}$$

If there is complete correlation, we see from Equation (7-4) that $bb' = 1$. If there is no correlation, both b and b' are 0. We therefore define the experimental linear-correlation coefficient $r \equiv \sqrt{bb'}$ as a measure of the degree of linear correlation.

$$r \equiv \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{[N\Sigma x_i^2 - (\Sigma x_i)^2]^{1/2}[N\Sigma y_i^2 - (\Sigma y_i)^2]^{1/2}} \tag{7-5}$$

The value of r ranges from 0, when there is no correlation, to ± 1, when there is complete correlation. The sign of r is the same as that of b (and b'), but only the absolute magnitude is important.

The correlation coefficient r cannot be used directly to indicate the degree of correlation. A probability distribution for r can be derived from the two-dimensional Gaussian distribution, but its evaluation requires a knowledge of the correlation coefficient ρ of the parent population. A more common test of r is to compare its value with the probability distribution for a parent population which is completely uncorrelated, that is, for which $\rho = 0$. Such a comparison will indicate whether or not it is probable that the data points could represent a sample derived from an uncorrelated parent population. If this probability is small, then it is more probable that the data points represent a sample from a parent population where the variables are correlated.

For a parent population with $\rho = 0$, the probability that any random sample of uncorrelated experimental data points would yield an experimental linear-correlation coefficient equal to r is given by[1]

$$P_r(r,\nu) = \frac{1}{\sqrt{\pi}} \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)} (1 - r^2)^{(\nu-2)/2} \tag{7-6}$$

where $\nu = N - 2$ is the number of degrees of freedom for an experimental sample of N data points.

The gamma function $\Gamma(n)$ is equivalent to the factorial function $n!$ extended to nonintegral arguments. It is defined for integral and half-integral arguments by the values for arguments of 1 and ½ and a recursion relation.

$$\Gamma(1) = 1 \qquad \Gamma(\tfrac{1}{2}) = \sqrt{\pi} \qquad \Gamma(n+1) = n\Gamma(n)$$

For integral arguments

$$\Gamma(n+1) = n! \qquad n = 0, 1, \ldots$$

For half-integral arguments

$$\Gamma(n+1) = n(n-1)(n-2) \cdots (\tfrac{3}{2})(\tfrac{1}{2}\sqrt{\pi}) \qquad n = \tfrac{1}{2}, \tfrac{3}{2}, \tfrac{5}{2}, \ldots \tag{7-7}$$

[1] For a derivation see Pugh and Winslow, sec. 12-8.

Integral probability A more useful distribution than that of Equation (7-6) is the probability $P_c(r,N)$ that a random sample of N uncorrelated experimental data points would yield an experimental linear-correlation coefficient as large as or larger than the observed value of $|r|$. This probability is the integral of $P_r(r,\nu)$ for $\nu = N - 2$.

$$P_c(r,N) = 2\int_{|r|}^{1} P_r(\rho,\nu)\, d\rho \qquad \nu = N - 2 \tag{7-8}$$

With this definition, $P_c(r,N)$ indicates the probability that the observed data could have come from an uncorrelated ($\rho = 0$) parent population. A small value of $P_c(r,N)$ implies that the observed variables are probably correlated.

Program 7-1 The probability function $P_c(r,N)$ of Equation (7-8) can be computed by expanding the integral. For even values of ν the exponent is an integer and the binomial expansion can be used to expand the argument of the integral.

$$P_c(r,N) = \begin{cases} 1 - \dfrac{2}{\sqrt{\pi}} \dfrac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)} \displaystyle\int_0^{|r|} \sum_{i=0}^{I} \left[(-1)^i \frac{I!}{(I-i)!i!} \rho^{2i} \right] d\rho & I = \tfrac{1}{2}(\nu-2) \\ 1 - \dfrac{2}{\sqrt{\pi}} \dfrac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)} \left\{ \displaystyle\sum_{i=0}^{I} \left[(-1)^i \frac{I!}{(I-i)!i!} \frac{|r|^{2i+1}}{2i+1} \right] \right\} & \nu \text{ even} \end{cases}$$

For odd values of ν, the exponent is half-integral and the expansion is more complex to derive, but the gamma functions may be included in the expansion to simplify the computation.

$$P_c(r,N) = 1 - \frac{1}{\sqrt{\pi}} \left\{ \sin^{-1}(|r|) + |r| \sum_{i=\frac{1}{2}}^{I} \left[(1-r^2)^i \frac{(2i-1)!!}{2i!!} \right] \right\} \qquad \nu \text{ odd}$$

Program 7-1 PCORRE Integral linear-correlation coefficient probability function $P_c(r,N)$.

```
C     FUNCTION PCORRE
C
C     PURPOSE
C       EVALUATE PROBABILITY FOR NO CORRELATION BETWEEN TWO VARIABLES
C
C     USAGE
C       RESULT = PCORRE (R, NPTS)
C
C     DESCRIPTION OF PARAMETERS
C       R      - LINEAR CORRELATION COEFFICIENT
C       NPTS   - NUMBER OF DATA POINTS
C
C     SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIRED
C       GAMMA (X)
C          CALCULATES GAMMA FUNCTION FOR INTEGERS AND HALF-INTEGERS
C
C     MODIFICATIONS FOR FORTRAN II
C       OMIT DOUBLE PRECISION SPECIFICATIONS
C       ADD F SUFFIX TO ABS IN STATEMENTS 23 AND 42
C       CHANGE DSQRT TO SQRTF IN STATEMENT 42
C       CHANGE DATAN TO ATANF IN STATEMENT 43
C
```

The double factorial sign !! represents

$$n!! = n(n-2)(n-4)\cdots\begin{matrix}(3)\ (1) \\ (4)\ (2)\end{matrix}\qquad\begin{matrix}\text{for } n \text{ odd} \\ \text{for } n \text{ even}\end{matrix}$$

The computation of $P_c(r,N)$ is illustrated in the computer routine PCORRE of Program 7-1. This is a Fortran function subprogram to evaluate $P_c(r,N)$ for a given value of r and N. The input variables are R $= r$, the correlation coefficient to be tested, and NPTS $= N$, the number of data points.

$$\text{FREE} = \text{NFREE} = \nu = N - 2$$

is the number of degrees of freedom for a linear fit, and IMAX $= I$ is the number of terms in the expansion. The sum of terms is accumulated in statements 31–36 for ν even and in statements 51–56 for ν odd. The value of the probability is returned to the calling program as the value of the function PCORRE.

Program 7-2 The computer routine GAMMA of Program 7-2 is used to evaluate the gamma functions. Statements 11–13 determine whether the argument of the calling sequence X is integral or half-integral. If the argument is integral, the gamma function

Program 7-1 PCORRE *(continued)*

```
      FUNCTION PCORRE (R, NPTS)
      DOUBLE PRECISION R2, TERM, SUM, FI, FNUM, DENOM
C
C        EVALUATE NUMBER OF DEGREES OF FREEDOM
C
   11 NFREE = NPTS - 2
      IF (NFREE) 13, 13, 15
   13 PCORRE = 0.
      GO TO 60
   15 R2 = R**2
      IF (1.-R2) 13, 13, 17
   17 NEVEN = 2*(NFREE/2)
      IF (NFREE - NEVEN) 21, 21, 41
C
C        NUMBER OF DEGREES OF FREEDOM IS EVEN
C
   21 IMAX = (NFREE-2)/2
      FREE = NFREE
   23 TERM = ABS (R)
      SUM = TERM
      IF (IMAX) 60, 26, 31
   26 PCORRE = 1. - TERM
      GO TO 60
   31 DO 36 I=1, IMAX
      FI = I
      FNUM = IMAX - I + 1
      DENOM = 2*I + 1
      TERM = -TERM * R2 * FNUM/FI
   36 SUM = SUM + TERM/DENOM
      PCORRE = 1.128379167 * (GAMMA((FREE+1.)/2.) / GAMMA(FREE/2.))
      PCORRE = 1. - PCORRE*SUM
      GO TO 60
C
C        NUMBER OF DEGREES OF FREEDOM IS ODD
C
   41 IMAX = (NFREE-3)/2
   42 TERM = ABS (R) * DSQRT(1.-R2)
   43 SUM = DATAN(R2/TERM)
      IF (IMAX) 57, 45, 51
   45 SUM = SUM + TERM
      GO TO 57
   51 SUM = SUM + TERM
   52 DO 56 I=1, IMAX
      FNUM = 2*I
      DENOM = 2*I + 1
      TERM = TERM * (1.-R2) * FNUM/DENOM
   56 SUM = SUM + TERM
   57 PCORRE = 1. - 0.6366197724*SUM
   60 RETURN
      END
```

is identical to the factorial function FACTOR(N) $= N!$ of Program 3-2, which is called in statement 21 to evaluate the result. If the argument is half-integral, the result GAMMA is set initially equal to $\Gamma(\frac{1}{2})$ in statement 31, and the product of Equations (7-7) is iterated in statements 41–43 for $x < 11$ and in statements 51–55 for $x > 10$.

Program 7-2 GAMMA Gamma function $\Gamma(n)$ for integers and half-integers.

```
C     FUNCTION GAMMA
C
C     PURPOSE
C       CALCULATE THE GAMMA FUNCTION FOR INTEGERS AND HALF-INTEGERS
C
C     USAGE
C       RESULT = GAMMA (X)
C
C     DESCRIPTION OF PARAMETERS
C       X      - INTEGER OR HALF-INTEGER
C
C     SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIRED
C       FACTOR (N)
C          CALCULATES N FACTORIAL FOR INTEGERS
C
C     MODIFICATIONS FOR FORTRAN II
C       OMIT DOUBLE PRECISION SPECIFICATIONS
C       CHANGE DLOG TO LOGF IN STATEMENT 54
C       CHANGE DEXP TO EXPF IN STATEMENT 55
C
      FUNCTION GAMMA (X)
      DOUBLE PRECISION PROD, SUM, FI
C
C        INTEGERIZE ARGUMENT
C
   11 N = X - .25
      XN = N
   13 IF (X-XN-.75) 31, 31, 21
C
C        ARGUMENT IS INTEGER
C
   21 GAMMA = FACTOR(N)
      GO TO 60
C
C        ARGUMENT IS HALF-INTEGER
C
   31 PROD = 1.77245385
      IF (N) 44, 44, 33
   33 IF (N-10) 41, 41, 51
   41 DO 43 I=1, N
      FI = I
   43 PROD = PROD * (FI-.5)
   44 GAMMA = PROD
      GO TO 60
   51 SUM = 0.
      DO 54 I=11, N
      FI = I
   54 SUM = SUM + DLOG(FI-.5)
   55 GAMMA = PROD * 639383.8623 * DEXP(SUM)
   60 RETURN
      END
```

Sample calculation The calculation of the linear-correlation coefficient R = r is carried out in the subroutine LINFIT of Program 6-1. Statement 71 is equivalent to Equation (7-5) with provision for including the standard deviations σ_i of the data points as weighting factors. Note that SUM = $\Sigma(1/\sigma_i^2)$ is substituted for N = NPTS, and DELTA = Δ is substituted for the left-hand term in the denominator of Equation (7-5). Each of the sums includes the proper weighting by σ_i^2 as determined by the variable MODE (see discussion of Section 6-3).

EXAMPLE 7-1 In the experiment of Example 6-1, the linear-correlation coefficient r is given by Equation (7-5) to be

$$r = \frac{9(2898) - 45(466.7)}{\sqrt{9(285) - (45^2)}\sqrt{9(29{,}828.65) - (466.7)^2}}$$
$$= \frac{26{,}082 - 21{,}002}{\sqrt{540}\sqrt{50649}} = 0.97$$

From the graph of Figure C-3, a value of $r = 0.97$ with $N = 9$ observations yields a probability of determining such a large correlation from an uncorrelated population as $P_c(r,N) < 0.001$. This means that it is extremely improbable that the variables T and x are linearly uncorrelated; i.e., the probability is high that they are correlated and that our fit to a straight line is justified.

Similarly, in the experiment of Example 6-2, the linear-correlation coefficient is given by Equation (7-5) with the weighting factor $\sigma_i^2 = y_i$ introduced.

$$r = \frac{.1919(675) - 16.911(10)}{\sqrt{.1919(1837.74) - (16.911)^2}\sqrt{.1919(652) - (10)^2}}$$
$$= \frac{129.53 - 169.11}{\sqrt{66.66}\sqrt{25.119}} = -0.97$$

Again, the probability $P_c(r,N)$ for $r = 0.97$ with $N = 10$ observations is less than 0.001 indicating that the change in the counting rate C is linearly correlated with time t with a high degree of probability.

7-2 CORRELATION BETWEEN MANY VARIABLES

If the dependent variable y is a function of more than one variable,

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots \qquad (7\text{-}9)$$

we might investigate the correlation between y and each of the independent variables x_j or we might also inquire into the possibility of correlation between the various different variables x_j if they are not independent.

To differentiate between the subscripts of Equations (7-5) and (7-9), let us use double subscripts on the variables x_{ij}. The first subscript i will represent the observation y_i as in the previous discussions. The second subscript j will represent the particular variable under investigation. Let us also rewrite Equation (7-5) for the linear-correlation coefficient r in terms of another quantity s_{jk}^2.

We define the *sample covariance* s_{jk}^2

$$s_{jk}^2 \equiv \frac{1}{N-1}\Sigma[(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)] \tag{7-10}$$

where the means $\bar{x}_j$ and $\bar{x}_k$ are given by

$$\bar{x}_j \equiv \frac{1}{N}\Sigma x_{ij} \qquad \text{and} \qquad \bar{x}_k \equiv \frac{1}{N}\Sigma x_{ik} \tag{7-11}$$

and the sums are taken over the range of the subscript i from 1 to N. With this definition, the sample variance for one variable s_j^2

$$s_j^2 \equiv s_{jj}^2 = \frac{1}{N-1}\Sigma(x_{ij} - \bar{x}_j)^2 \tag{7-12}$$

is analogous to the sample variance s_x^2 defined in Equation (2-10).

$$s_x^2 = \frac{1}{N-1}\Sigma(x_i - \bar{x})^2$$

Equation (7-10) can be rewritten for comparison with Equation (7-5) by substituting the definitions of Equations (7-11).

$$\begin{aligned} s_{jk}^2 &= \frac{1}{N-1}\Sigma(x_{ij}x_{ik} - \bar{x}_j\bar{x}_k) \\ &= \frac{1}{N-1}\left(\Sigma x_{ij}x_{ik} - \frac{1}{N}\Sigma x_{ij}\Sigma x_{ik}\right) \end{aligned} \tag{7-13}$$

If we substitute x_{ij} for x_i and x_{ik} for y_i in Equation (7-5), we can define the *sample linear-correlation coefficient* between any two variables x_j and x_k as

$$r_{jk} \equiv \frac{s_{jk}^2}{s_j s_k} \tag{7-14}$$

with the covariances and variances s_{jk}^2, s_j^2, and s_k^2 given by Equations (7-12) and (7-13). Thus, the linear-correlation coefficient between the jth variable x_j and the dependent variable y is given by

$$r_{jy} = \frac{s_{jy}^2}{s_j s_y} \tag{7-15}$$

Similarly, the linear-correlation coefficient of the parent population of which the data are a sample is defined as

$$\rho_{jk} = \frac{\sigma_{jk}^2}{\sigma_j \sigma_k}$$

where σ_j^2, σ_k^2, and σ_{jk}^2 are the true variances and covariances of the parent population. These linear-correlation coefficients are also known as product-moment-correlation coefficients.

With such a definition we can consider either the correlation between the dependent variable and any other variable r_{jy} or the correlation between any two variables r_{jk}. It is important to note, however, that the sample variances s_j^2 defined by Equation (7-12) are measures of the range of variation of the variables and not of the uncertainties as are the sample variances s^2 defined in Sections 5-2 and 6-5.

Polynomials In Chapter 8 we will investigate functional relationships between y and x of the form

$$y = a + bx + cx^2 + dx^3 + \cdots \tag{7-16}$$

In a sense, this is a variation on the linear relationship of Equation (7-8) where the powers of the single independent variable x are considered to be various variables $x_j = x^j$. The correlation between the independent variable y and the mth term in the

power series of Equation (7-16), therefore, can be expressed in terms of Equations (7-12)–(7-15).

$$r_{my} = \frac{s_{my}^2}{s_m s_y}$$

$$s_m^2 = \frac{1}{N-1}\left[\Sigma x_i^{2m} - \frac{1}{N}(\Sigma x_i^m)^2\right]$$

$$s_y^2 = \frac{1}{N-1}\left[\Sigma y_i^2 - \frac{1}{N}(\Sigma y_i)^2\right]$$

$$s_{my}^2 = \frac{1}{N-1}\left(\Sigma x_i^m y_i - \frac{1}{N}\Sigma x_i^m \Sigma y_i\right)$$

Weighted fit If the uncertainties of the data points are not all equal $\sigma_i \neq \sigma$, we must include the individual standard deviations σ_i as weighting factors in the definitions of variances, covariances, and correlation coefficients. From Section 6-3, the prescription for introducing weighting is to weight each term in a sum by the factor $1/\sigma_i^2$.

The formula for the correlation coefficient remains the same as Equations (7-14) and (7-15), but the formulas of Equations (7-10) and (7-12) for calculating the variances and covariances must be modified

$$s_{jk}^2 \equiv \frac{\frac{1}{N-1}\sum\left[\frac{1}{\sigma_i^2}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\right]}{\frac{1}{N}\sum\frac{1}{\sigma_i^2}}$$

$$s_j^2 \equiv s_{jj}^2 = \frac{\frac{1}{N-1}\sum\left[\frac{1}{\sigma_i^2}(x_{ij} - \bar{x}_j)^2\right]}{\frac{1}{N}\sum\frac{1}{\sigma_i^2}} \qquad (7\text{-}17)$$

where the means $\bar{x}_j$ and $\bar{x}_k$ are also weighted means.

$$\bar{x}_j \equiv \frac{\Sigma[(1/\sigma_i^2)x_{ij}]}{\Sigma(1/\sigma_i^2)}$$

Thus, the actual weighting factor is

$$\text{Weight}_i = \frac{1/\sigma_i^2}{(1/N)\Sigma(1/\sigma_i^2)}$$

as specified by the discussion of Chapter 4 and Section 10-1.

Multiple-correlation coefficient We can extrapolate the concept of the linear-correlation coefficient, which characterizes the correlation between two variables at a time, to include multiple correlations between groups of variables taken simultaneously.

The linear-correlation coefficient r of Equation (7-5) between y and x can be expressed in terms of the variances and covariances of Equations (7-17) and the slope b of a straight-line fit given in Equation (7-2).

$$r^2 = \frac{s_{xy}^4}{s_x^2 s_y^2} = b\frac{s_{xy}^2}{s_y^2}$$

In analogy with this definition of the linear-correlation coefficient, we define the *multiple-correlation coefficient* R to be the sum over similar terms for the variables of Equation (7-9).

$$R^2 \equiv \sum_{j=1}^{n}\left(b_j\frac{s_{jy}^2}{s_y^2}\right) = \sum_{j=1}^{n}\left(b_j\frac{s_j}{s_y}r_{jy}\right) \qquad (7\text{-}18)$$

The linear-correlation coefficient r is useful for testing whether one particular variable should be included in the theoretical function to which the data are fit. The multiple-correlation coefficient R characterizes the fit of the data to the entire function. A comparison of the multiple-correlation coefficient for different functions is therefore useful in optimizing the theoretical functional form.

We will defer until Chapter 10 a discussion of how to use these correlation coefficients to determine the validity of including each term in the polynomial of Equation (7-16) or the arbitrary function of Equation (7-9).

SUMMARY

Function linear in coefficients:

$$y = a + \sum_{j=1}^{n} b_j x_j$$

Sample covariance s_{jk}^2:

$$s_{jk}^2 \equiv \frac{\frac{1}{N-1}\sum_i \left[\frac{1}{\sigma_i^2}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\right]}{\frac{1}{N}\sum_i \frac{1}{\sigma_i^2}}$$

$$= \frac{1}{N-1}\left(\frac{\sum \frac{x_{ij}x_{ik}}{\sigma_i^2}}{\frac{1}{N}\sum \frac{1}{\sigma_i^2}} - N\bar{x}_j\bar{x}_k\right)$$

$$\bar{x}_j \equiv \frac{\Sigma(x_{ij}/\sigma_i^2)}{\Sigma(1/\sigma_i^2)}$$

Sample variance:

$$s_j^2 \equiv s_{jj}^2$$

Linear-correlation coefficient:

$$r_{jk} \equiv \frac{s_{jk}^2}{s_j s_k}$$

Probability $P_c(r,N)$ that any random sample of uncorrelated experimental data points would yield an experimental linear-correlation coefficient as large as or larger than $|r|$:

$$P_c(r, \nu + 2) = \int_{|r|}^{1} \frac{2}{\sqrt{\pi}} \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)} (1 - x^2)^{(\nu-1)/2}\, dx$$

Multiple-correlation coefficient R:

$$R^2 \equiv \sum_{j=1}^{n} \left(b_j \frac{s_{jy}^2}{s_y^2}\right) = \sum_{j=1}^{n} \left(b_j \frac{s_j}{s_y} r_{jy}\right)$$

EXERCISES

7-1 Find the linear-correlation coefficient r for Example 6-1.

7-2 If a set of data when fitted with Equation (7-1) yield a zero slope $b = 0$, what can you say about the linear-correlation coefficient r? Justify this value in terms of the correlation between x_i and y_i.

7-3 Find the linear-correlation coefficient r for Example 6-2.

7-4 Verify the expansion in the computation of $P_c(r,N)$.

7-5 Find the linear-correlation coefficient r_1 between x_i and y_i for the data of Example 8-1.

7-6 If the linear-correlation coefficient r_1 of Exercise 7-5 is computed by evaluating the slopes b and b' ($r_1 = \sqrt{bb'}$), should the slopes be computed by fitting the data with a linear polynomial or a quadratic polynomial?

7-7 Find the correlation coefficient r_2 between x_i^2 and y_i for the data of Example 8-1. Does the correlation justify the use of a quadratic term?

7-8 Express the multiple-correlation coefficient R in terms of x_{ij}, y_i, and their averages.

7-9 Evaluate the multiple-correlation coefficient R for the data of Example 8-1.