USGS
science for a changing world

Techniques of Water-Resources Investigations
of the United States Geological Survey

Chapter B4

REGRESSION MODELING OF GROUND-WATER FLOW

By Richard L. Cooley and Richard L. Naff

Book 3
APPLICATIONS OF HYDRAULICS

UNITED STATES DEPARTMENT OF THE INTERIOR

MANUEL LUJAN, JR., Secretary

GEOLOGICAL SURVEY

Dallas L. Peck, Director

UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON : 1990

For sale by the Books and Open-File Reports Section, U.S. Geological Survey,
Federal Center, Box 25425, Denver, CO 80225

Preface

Scientists and engineers have been using ground-water flow models to study ground-water flow systems for more than 20 years. The basic modeling process seems to be relatively straightforward. Initially, a sound conceptual model is formed and is translated into a tractable, mathematical model. Contributing to (and following) this conceptualization process is the collection of field information, such as (1) location and extent of hydrostratigraphic units, recharge areas, discharge areas, and system boundaries; (2) hydraulic head measurements; and (3) pumping discharges. These data form the basis for input to the flow model. Finally, the model is run, and the desired information such as head distribution or flux rates is extracted. However, people engaged in modeling usually observe that two pervasive problems considerably complicate the situation. One problem is that good, general methods of measuring (or computing) some of the variables that characterize the flow system and its geologic framework do not exist. One example is measurement of ground-water recharge. No direct ways of measuring recharge exist, and the accuracy of indirect methods is often unknown. Furthermore, many indirect methods are applicable only to unique situations. The second problem relates to errors in the measurements and their propagation into model results. No error-free measurement (or computation) methods for obtaining data on the flow system exist. Thus, even the variables that can be estimated will contribute to error, so that model results will always be unreliable to some extent. As a consequence of these two problems, measurement (or computation) of the necessary input variables, application of them to an adequate model, and calculation of the desired results to an acceptable accuracy generally are not possible. Other methods that recognize and deal with the problems of incompleteness and (or) inaccuracy of data must also be applied. The present text has been designed to teach these methods to scientists and engineers engaged in ground-water modeling.

The basic methodology is multiple, nonlinear regression, in which the regression model is some type of ground-water flow model. As seen subsequently, this methodology is consistent with known aspects of the physical systems to be analyzed and requires relatively few assumptions. Even though the present text is directed specifically toward ground-water modeling, the procedures to be discussed are applicable to a number of different types of modeling problems. Thus, the methods are usually discussed in a general context; in other words, without reference to any specific model.

Material in the present text evolved from notes developed for training courses in parameter estimation for ground-water flow models taught by the authors and others at the U.S. Geological Survey National Training Center, Denver Federal Center, Lakewood, Colo. The philosophy of these courses, and of this text, is to teach general methods that are applicable to a wide range of problems and to teach these methods in sufficient depth so that students can apply them to many problem situations not considered in the courses or text.

The main body of the text is organized into six major sections. The first section is an introduction that discusses the general topic of modeling ground-water flow. This section shows that ground-water modeling problems are an incomplete combination of direct-type problems (solution for hydraulic head given values of flow system and framework variables) and inverse-type problems (solution for flow system and framework variables given values of hydraulic head) that commonly require solution by optimization procedures which give the best fit between observed and calculated results. Because the specific optimization approach employed here is regression and regression procedures are based on statistical concepts, the second section is included to provide the student with the necessary statistical background material. It is not designed to be an exhaustive review of basic statistics; rather, it presents material essential to understanding the following sections. The third section presents detailed material on linear and nonlinear regression. Although most

of the material on linear regression is fairly standard, some of the material on nonlinear regression is not. In particular, specific modifications presented to induce convergence of the iterative solution procedure for nonlinear regression have not, to the writers' knowledge, been presented elsewhere in the form given here. The fourth section applies the nonlinear regression method to the specific problem of developing a general finite-difference model of steady-state ground-water flow. In the fifth section, statistical procedures are given to analyze and use general linear and nonlinear regression models. The tests and analytical procedures presented are not exhaustive; they are the ones that the writers have found to be most useful for analyzing the real systems examined to date. The sixth section is designed to be supplemental to the preceding sections. Specialized procedures presented include nonlinear regression for models that cannot be solved directly for the dependent variable, a measure of model nonlinearity called Beale's measure, and a statistical test for compatibility of prior information on parameters and parameter estimates derived from sample (observed head) information.

A number of exercises have been included, and a complete discussion of the answers can be found in the seventh major section at the end of the text. These problems exercise the student on nearly all methods presented. In addition, three computer programs are documented and listed: the program for nonlinear-regression solution of ground-water flow problems of section four, a program to calculate Beale's measure, and a program to calculate simulated errors in computed dependent variables such as hydraulic head.

The mathematical background necessary to use this text includes basic mathematics through differential and integral calculus, including partial derivatives, and matrix algebra. A background in elementary statistics would be useful but is not essential. In addition, a sound knowledge of ground-water hydrology and ground-water flow modeling are needed to effectively apply the methods presented.

References for cited material are given at the end of each major section. Good supplemental sources for the unreferenced material not peculiar to this text are presented as "Additional Reading" at the end of each reference list. It is expected that students who have difficulty with the material in this text will consult the more expanded developments in these supplemental sources.

Several people, in addition to the writers, contributed extensively to this text. Charles R. Faust wrote earlier sections on statistical review and basic regression and contributed several exercises, Steven P. Larson wrote an earlier version and documentation of the nonlinear-regression flow program of section four and contributed earlier versions of several exercises, James V. Tracy contributed to the documentation of the nonlinear-regression flow program, and Thomas Maddock III wrote the first version of the statistics review section. In addition, all of these people helped teach the training courses from which the present text evolved. Finally, the writers would like to thank the technical reviewers, Brent M. Troutman and Allan L. Gutjahr, for their many hours of review work and the secretaries, Anita Egelhoff, Evelyn R. Warren, and Patricia A. Griffith, for their patience and care in typing the manuscript.

TECHNIQUES OF WATER-RESOURCES INVESTIGATIONS OF THE UNITED STATES GEOLOGICAL SURVEY

The U.S. Geological Survey publishes a series of manuals describing procedures for planning and conducting specialized work in water-resources investigations. The manuals published to date are listed below and may be ordered by mail from the **U.S. Geological Survey, Books and Open-File Reports, Federal Center, Box 25425, Denver, Colorado 80225** (an authorized agent of the Superintendent of Documents, Government Printing Office).

Prepayment is required. Remittance should be sent by check or money order payable to U.S. Geological Survey. Prices are not included in the listing below as they are subject to change. **Current prices can be obtained** by writing to the USGS, Books and Open File Reports. Prices include cost of domestic surface transportation. For transmittal outside the U.S.A. (except to Canada and Mexico) a surcharge of 25 percent of the net bill should be included to cover surface transportation. When ordering any of these publications, please give the title, book number, chapter number, and "U.S. Geological Survey Techniques of Water-Resources Investigations."

TWI 1-D1. Water temperature—influential factors, field measurement, and data presentation, by H.H. Stevens, Jr., J.F. Ficke, and G.F. Smoot. 1975. 65 pages.

TWI 1-D2. Guidelines for collection and field analysis of ground-water samples for selected unstable constituents, by W.W. Wood. 1976. 24 pages.

TWI 2-D1. Application of surface geophysics to ground-water investigations, by A.A.R. Zohdy, G.P. Eaton, and D.R. Mabey. 1974. 116 pages.

TWI 2-E1. Application of borehole geophysics to water-resources investigations, by W.S. Keys and L.M. MacCary. 1971. 126 pages.

TWI 3-A1. General field and office procedures for indirect discharge measurement, by M.A. Benson and Tate Dalrymple. 1967. 30 pages.

TWI 3-A2. Measurement of peak discharge by the slope-area method, by Tate Dalrymple and M.A. Benson. 1967. 12 pages.

TWI 3-A3. Measurement of peak discharge at culverts by indirect methods, by G.L. Bodhaine. 1968. 60 pages.

TWI 3-A4. Measurement of peak discharge at width contractions by indirect methods, by H.F. Matthai. 1967. 44 pages.

TWI 3-A5. Measurement of peak discharge at dams by indirect methods, by Harry Hulsing. 1967. 29 pages.

TWI 3-A6. General procedure for gaging streams, by R.W. Carter and Jacob Davidian. 1968. 13 pages.

TWI 3-A7. Stage measurements at gaging stations, by T.J. Buchanan and W.P. Somers. 1968. 28 pages.

TWI 3-A8. Discharge measurements at gaging stations, by T.J. Buchanan and W.P. Somers. 1969. 65 pages.

TWI 3-A9. Measurement of time of travel and dispersion in streams by dye tracing, by E.P. Hubbard, F.A. Kilpatrick, L.A. Martens, and J.F. Wilson, Jr. 1982. 44 pages.

TWI 3-A10. Discharge ratings at gaging stations, by E.J. Kennedy. 1984. 59 pages.

TWI 3-A11. Measurement of discharge by moving-boat method, by G.F. Smoot and C.C. Novak. 1969. 22 pages.

TWI 3-A12. Fluorometric procedures for dye tracing, Revised, by James F. Wilson, Jr., Ernest D. Cobb, and Frederick A. Kilpatrick. 1986. 41 pages.

TWI 3-A13. Computation of continuous records of streamflow, by Edward J. Kennedy. 1983. 53 pages.

TWI 3-A14. Use of flumes in measuring discharge, by F.A. Kilpatrick, and V.R. Schneider. 1983. 46 pages.

TWI 3-A15. Computation of water-surface profiles in open channels, by Jacob Davidian. 1984. 48 pages.

TWI 3-A16. Measurement of discharge using tracers, by F.A. Kilpatrick and E.D. Cobb. 1985. 52 pages.

TWI 3-A17. Acoustic velocity meter systems, by Antonius Laenen. 1985. 38 pages.

TWI 3-B1. Aquifer-test design, observation, and data analysis, by R.W. Stallman. 1971. 26 pages.

TWI 3-B2. Introduction to ground-water hydraulics, a programmed text for self-instruction, by G.D. Bennett. 1976. 172 pages. Spanish translation TWI 3-B2 also available.

TWI 3-B3. Type curves for selected problems of flow to wells in confined aquifers, by J.E. Reed. 1980. 106 p.

TWI 3-B4. Regression modeling of ground-water flow, by Richard L. Cooley and Richard L. Naff. 1989. 232 pages.

TWI 3-B5. Definition of boundary and initial conditions in the analysis of saturated ground-water flow systems—an introduction, by O. Lehn Franke, Thomas E. Reilly, and Gordon D. Bennett. 1987. 15 pages.

TWI 3-B6. The principle of superposition and its application in ground-water hydraulics, by Thomas E. Reilly, O. Lehn Franke, and Gordon D. Bennett. 1987. 28 pages.

TWI 3-C1. Fluvial sediment concepts, by H.P. Guy. 1970. 55 pages.
TWI 3-C2. Field methods of measurement of fluvial sediment, by H.P. Guy and V.W. Norman. 1970. 59 pages.
TWI 3-C3. Computation of fluvial-sediment discharge, by George Porterfield. 1972. 66 pages.
TWI 4-A1. Some statistical tools in hydrology, by H.C. Riggs. 1968. 39 pages.
TWI 4-A2. Frequency curves, by H.C. Riggs, 1968. 15 pages.
TWI 4-B1. Low-flow investigations, by H.C. Riggs. 1972. 18 pages.
TWI 4-B2. Storage analyses for water supply, by H.C. Riggs and C.H. Hardison. 1973. 20 pages.
TWI 4-B3. Regional analyses of streamflow characteristics, by H.C. Riggs. 1973. 15 pages.
TWI 4-D1. Computation of rate and volume of stream depletion by wells, by C.T. Jenkins. 1970. 17 pages.
TWI 5-A1. Methods for determination of inorganic substances in water and fluvial sediments, by M.W. Skougstad and others, editors. 1979. 626 pages.
TWI 5-A2. Determination of minor elements in water by emission spectroscopy, by P.R. Barnett and E.C. Mallory, Jr. 1971. 31 pages.
TWI 5-A3. Methods for the determination of organic substances in water and fluvial sediments, edited by R.L. Wershaw, M.J. Fishman, R.R. Grabbe, and L.E. Lowe. 1987. 80 pages. This manual is a revision of "Methods for Analysis of Organic Substances in Water" by Donald F. Goerlitz and Eugene Brown, Book 5, Chapter A3, published in 1972.
TWI 5-A4. Methods for collection and analysis of aquatic biological and microbiological samples, edited by P.E. Greeson, T.A. Ehlke, G.A. Irwin, B.W. Lium, and K.C. Slack. 1977. 332 pages.
TWI 5-A5. Methods for determination of radioactive substances in water and fluvial sediments, by L.L. Thatcher, V.J. Janzer, and K.W. Edwards. 1977. 95 pages.
TWI 5-A6. Quality assurance practices for the chemical and biological analyses of water and fluvial sediments, by L.C. Friedman and D.E. Erdmann. 1982. 181 pages.
TWI 5-C1. Laboratory theory and methods for sediment analysis, by H.P. Guy. 1969. 58 pages.
TWI 6-A1. A modular three-dimentional finite-difference ground-water flow model, by Michael G. McDonald and Arlen W. Harbaugh. 1988. 586 pages.
TWI 7-C1. Finite difference model for aquifer simulation in two dimensions with results of numerical experiments, by P.C. Trescott, G.F. Pinder, and S.P. Larson. 1976. 116 pages.
TWI 7-C2. Computer model of two-dimensional solute transport and dispersion in ground water, by L.F. Konikow and J.D. Bredehoeft. 1978. 90 pages.
TWI 7-C3. A model for simulation of flow in singular and interconnected channels, by R.W. Schaffranek, R.A. Baltzer, and D.E. Goldberg. 1981. 110 pages.
TWI 8-A1. Methods of measuring water levels in deep wells, by M.S. Garber and F.C. Koopman. 1968. 23 pages.
TWI 8-A2. Installation and service manual for U.S. Geological Survey monometers, by J.D. Craig. 1983. 57 pages.
TWI 8-B2. Calibration and maintenance of vertical-axis type current meters, by G.F. Smoot and C.E. Novak. 1968. 15 pages.

CONTENTS

COMPUTER DISKETTE INFORMATION

This report contains a computer diskette (in the pocket at the back of the report) that has the source codes and data sets to be used with the report. The source codes were developed using the Microsoft Fortran Compiler, Version 3.3, with the DOS 2.0 operating system on an IBM PC/XT computer with the IBM 8088 Math Coprocessor and 256KB memory. Except for the OPEN statements near the beginning of the codes, Fortran 66 was used throughout to make the codes as machine independent as possible. For more information concerning the contents of the diskette, insert the diskette in drive A of the computer, type README, and push the enter key.

METRIC CONVERSION FACTORS

For those readers who prefer to use metric units, conversion factors for terms used in this report are listed below:

Multiply	*By*	*To obtain*
foot (ft)	0.3048	meter (m)
square foot (ft^2)	0.09290	square meter (m^2)
foot per day (ft/d)	0.3048	meter per day (m/d)
square foot per second (ft^2/s)	0.09290	square meter per second (m^2/s)
square foot per day (ft^2/d)	0.09290	square meter per day (m^2/d)
gallon per day per foot (gal/d/ft)	0.01242	square meter per day (m^2/d)
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)
cubic foot per day (ft^3/d)	0.02832	cubic meter per day (m^2/d)
gallon per minute (gal/min)	0.00006309	cubic meter per second (m^3/s)
ounces per ton	31.25	grams per metric ton[1]

[1] 1 metric ton = 1 megagram

REGRESSION MODELING OF GROUND-WATER FLOW

By Richard L. Cooley and Richard L. Naff

1 Introduction

1.1 Flow Equation and Boundary Conditions

The most general form of the ground-water flow equation that we consider here is given as

$$\frac{\partial}{\partial x}\left(T_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(T_{yy}\frac{\partial h}{\partial y}\right)+R(H-h)+W$$

$$+\sum_{\ell=1}^{N}\delta(x-a_\ell)\delta(y-b_\ell)Q_\ell=S\frac{\partial h}{\partial t} \qquad (1.1\text{-}1)$$

where

$T_{\xi\xi}(x,y)$= transmissivity ($K_{\xi\xi}b$) in the $\xi=x$ or y direction;
$K_{\xi\xi}(x,y)$= hydraulic conductivity of the aquifer in the ξ direction;
$b(x,y)$= thickness of the aquifer;
$R(x,y)$= hydraulic conductance (hydraulic conductivity divided by thickness) of sediments underlying a stream or of an aquitard underlying or overlying the aquifer;
$W(x,y,t)$= source-sink term (positive for a source), distributed areally;
$\sum_{\ell=1}^{N}\delta(x-a_\ell)\delta(y-b_\ell)Q_\ell$=Dirac delta designation for N wells, each one pumping at rate $Q_\ell(t)$ (positive for injection) and located at (a_ℓ,b_ℓ);
$S(x,y)$= storage coefficient;
$h(x,y,t)$= hydraulic head in the aquifer;
$H(x,y,t)$= head at the stream bottom or at the distal side of the aquitard;
x,y= Cartesian coordinates;
t= time;

and T_{xx} and T_{yy} are continuous functions of x and y.

With suitable internal boundary conditions, the region can be zoned with respect to $T_{\xi\xi}$. Such boundary conditions involve head and specific discharge multiplied by thickness normal to the boundary (q_n) and can be stated for a boundary between $T_{\xi\xi}$ zones k and ℓ as

$$(h)_k=(h)_\ell \qquad (1.1\text{-}2)$$

$$(q_n)_k=(q_n)_\ell \qquad (1.1\text{-}3)$$

where $(\cdot)_k$ indicates that the quantity in parentheses is evaluated just within the k side of the boundary and similarly for ℓ. Zonation with respect to R, W, or S requires no internal boundary conditions.

External boundary conditions applying at the periphery of the domain being modeled are given as

$$T_{xx}\frac{\partial h}{\partial x}n_x+T_{yy}\frac{\partial h}{\partial y}n_y+\alpha h=\beta \qquad (1.1\text{-}4)$$

where $\alpha(x,y,t)$ and $\beta(x,y,t)$ are given functions, and $\{n_x(x,y),n_y(x,y)\}$ is the outward-pointing unit normal at the boundary. The sum of the first two terms is the flux q_B normal to the boundary (positive for outflow). Equation 1.1-4

incorporates the standard boundary conditions of specified flux (q_B) and specified head (h_B) but also allows for linear combinations to be given.

1.2 Types of Solutions

1.2.1 Direct Solution for Head

The classical problem of mathematical physics (and, by assumption, of ground-water hydrology) is to directly solve equations 1.1-1 through 1.1-4 for $h=h(x,y,t)$. Given that any specific problem is properly posed, such a solution will always exist. The conditions for properly posing a problem are the following.

1. The positions of all internal boundaries are known exactly. Examples of internal boundaries are abrupt changes in $T_{\xi\xi}$, R, S, or W; internal known flux (q_n) boundaries; and internal known head boundaries. Note that a river is often treated as either an internal known head boundary where the river is assumed to have no width, or a zone of differing R where each bank is a zone boundary.
2. The positions and types of all external boundaries are known exactly. External boundaries frequently are known flux (q_B) types or known head (h_B) types. Sometimes some linear combination is known.
3. Hydrogeologic variables $T_{\xi\xi}$, R, and S and hydrologic variables W and Q_ℓ are known at all points in the region.
4. All boundary-condition variables H, α, and β are known. The initial head (at $t=0$) is a boundary condition and must also be known.

Obviously, ground-water flow problems are not actually of the classical type because none of the conditions cited above ever are met exactly. Conditions 1 and 2 are often most closely fulfilled, but estimates (often crude) usually must suffice for the variables in conditions 3 and 4. Any errors in these input variables are propagated directly into the solution. However, reasonable (but incorrect) estimates of the variables can be shown to yield errors in predicted $h(x,y,t)$ that have the characteristic of being bounded (that is, they do not tend to plus or minus infinity). Also, as the errors in the input variables tend to zero, the errors in computed head do also.

1.2.2 Inverse Solution for Parameters

An inverse solution involves solving equations 1.1-1 through 1.1-4 for one or more of the variables $T_{\xi\xi}$, R, S, W, Q_ℓ, α, or β, over the region; these variables are termed parameters here. Because R, S, W, Q_ℓ, α, and β are not involved in derivatives, theoretically they may be solved for algebraically. Unless $T_{\xi\xi}$ is constant, it is involved in derivatives and, thus, must be obtained by solving a differential equation. To understand this, note that equation 1.1-1 may be rearranged to give

$$a\frac{\partial T_{xx}}{\partial x}+b\frac{\partial T_{yy}}{\partial y}+cT_{xx}+dT_{yy}-F=0 \quad (1.2\text{-}1)$$

where

$$a(x,y,t)=\frac{\partial h}{\partial x}\ , \quad b(x,y,t)=\frac{\partial h}{\partial y}\ ,$$

$$c(x,y,t)=\frac{\partial^2 h}{\partial x^2}\ , \quad d(x,y,t)=\frac{\partial^2 h}{\partial y^2}\ , \text{ and}$$

$$F(x,y,t)=S\frac{\partial h}{\partial t}-R(H-h)-W-\sum_{\ell=1}^{N}\delta(x-a_\ell)\delta(y-b_\ell)Q_\ell .$$

In general, if T_{xx} and T_{yy} are known, conditions for finding R, S, W, Q_ℓ, α, or β are:

1. Conditions 1 and 2 for the classical direct solution are met.
2. Head distribution $h(x,y,t)$ is known exactly.
3. The solution for the desired combination of parameters to be obtained is unique.

The latter condition is completely problem dependent. Because solution involves only algebraic manipulations, the condition reduces to the requirement that the system of algebraic equations involving the desired parameters has a unique solution. Generally, solution involves picking the required number of points spatially and through time to yield the necessary number of equations.

To find T_{xx} and T_{yy}, more conditions are

required than for finding R, S, W, Q_p, α, or β. These conditions are:

1. Conditions 1 and 2 for finding R, S, W, Q_p, α, or β must be met.
2. The direction of the velocity vector must be known everywhere, or T_{xx}/T_{yy} must be known everywhere, or quantities a, b, c, d, and F in equation 1.2-1 must be known at two (or more) points in time to give a unique solution to equation 1.2-1 written in the form of a pair of simultaneous differential equations. These requirements result because 1.2-1 is one equation in two unknowns. Hence, an additional relationship is required. If the velocity direction is known everywhere, then by employing Darcy's law the additional relationship is derived as

$$\frac{T_{xx}}{T_{yy}} = \frac{q_x b}{q_y a} \qquad (1.2\text{-}2)$$

   where a, b, and q_x/q_y (the ratio of the x and y direction fluxes) are known.
3. If either the direction of the velocity vector or T_{xx}/T_{yy} is known, then either T_{xx} or T_{yy} must be known on a possibly discontinuous curve crossing all flowlines. If solution is to be obtained by solving a simultaneous pair of differential equations, then T_{xx} must be known on a possibly discontinuous curve that spans the range of y, and T_{yy} must be known on a possibly discontinuous curve that spans the range of x. These are extensions of the Cauchy boundary condition for a first-order differential equation involving a single dependent variable and are required for solution of the problem.
4. The function F in equation 1.2-1 must be known everywhere. This means that all quantities in F must be known or that a mathematical form for F can be assumed.

Ground-water flow modeling does not fit into the category of inverse solutions, although a significant part of most model studies is to find values of the parameters that allow values of calculated head to match those observed in the field. The difficulty is that the required conditions are almost never met. Head distribution is never known exactly because measurements do not exist at all points and, where these measurements do exist, they are not exact. Furthermore, some measure of $T_{\xi\xi}$ is virtually never available on the required curves, and information on directions of flow vectors for even scattered locations usually is nonexistent. Assumptions concerning zonations in which T_{xx}/T_{yy} and (or) T_{xx} and T_{yy} may be considered constant simplify the problem, but the fact that h must be known still remains.

Because the head distribution is not known exactly, coefficients a, b, c, d, and F in equation 1.2-1 are in error. Furthermore, head appears as a derivative in all of these quantities. Hence, any error in h is propagated into the inverse solution as a derivative of error. The effects of this propagation are often disastrous because, if ϵ_h is defined as error in head, $\epsilon_h \to 0$ does not imply that $\partial\epsilon_h/\partial\xi \to 0$. Also, $|\partial\epsilon_h/\partial\xi| >> |\epsilon_h|$ is common, and it can happen that $|\partial\epsilon_h/\partial\xi| \to \infty$ even if ϵ_h is bounded. Therefore, the error in computed $T_{\xi\xi}$ (or other parameter) may not approach zero as $\epsilon_h \to 0$, and may, in fact, be quite large (Neuman, 1980, p. 342-344).

1.2.3 Solution Using Real Data

In the previous section, we argued that problems involving ground-water flow modeling of real field systems are neither of the classical nor inverse type, because the data necessary for the problems to be classified as either type are usually lacking. An estimate of the hydraulic head distribution based on measurements (that are in error with respect to the model) taken at selected points usually exist. Estimates of the parameters are usually either completely unknown or have been obtained by spot measurements, few of which are directly useful for construction of appropriate effective values for use in equation 1.1-1. That modeling problems in ground-water hydrology involve an incomplete combination of several types of data in which error and error propagation are important considerations is evident.

1.3 Sources of Error in Ground-Water Data

Uncertainty (or errors) in ground-water data may have many sources, and enumeration of all

possible sources would be a nearly impossible task. However, a consideration of some of the more important sources of error serves to illustrate the importance of the error component.

1.3.1 Sources of Error in Head Data

Some major potential sources of random-appearing error in head data with respect to the model (equations 1.1–1 through 1.1–4) are:

1. Areal ground-water models assume that the head used is the average over the vertical. However, wells may not be open over the entire interval modeled, and if they are, they may not measure the average. Flow into and (or) out of a well distorts the hydraulic head field in the vicinity of the well so that the recorded water level does not represent the average head.
2. Permeability varies from point to point, which causes water levels to vary from values they would have if permeability were uniform. However, models usually do not take this detailed variation into account. This phenomenon has been extensively studied during the last 10 years, and literature reviews are contained in the works of Dagan (1986) and Gelhar (1984, 1986).
3. Water levels measured in wells in use may contain unknown amounts of residual drawdown. In addition, unused wells may be near wells that are in use, with resulting unknown drawdown in the unused well.
4. Measurement of well-head elevation may be in error.

Actual total error from the above sources is highly problem dependent, but it is easy to imagine errors of several feet. It should be noted that measurement error in water levels was not mentioned as a major source of error because it commonly amounts to one- or two-tenths of one foot or less. Finally, major model error in equations 1.1–1 through 1.1–4 (for example, head dependence in one or more parameters or three-dimensional flow) was also not mentioned because error resulting from this source is bias and should be detected and eliminated by analysis of model results.

1.3.2 Sources of Error in Parameter Data

Because there are several different parameters to be considered, and each can be estimated or measured in several different ways, a large number of sources of error exist in parameter data. Model error is not considered here, but other types of bias are potentially important and are often difficult to detect. Some examples of errors in parameter data illustrating the nature of the problem are:

1. Too few estimates of parameters are available to compute stable estimates of statistics, such as mean and variance.
2. Results of point sampling are often biased because a large amount of data does not necessarily allow computation of nearly true or effective values of a parameter and its variance. For example, permeability values from core analyses often are not representative of regional values, because flow through large fractures is not reproduced by core analyses. Also, effective values of a parameter and its variability are usually not directly given by standard mean and variance formulas.
3. Transmissivities estimated from specific-capacity data collected by drillers are subject to numerous sources of error. Common sources include (1) mismeasuring water levels or pumping rates, (2) allowing the water level to recover after bailing, (3) clogging of the slots or screen, and (4) inaccurate reporting. There are so many sources of error that the errors may often appear to be random. A persistent source of bias results because drillers drill wells in favorable locations and only screen (or slot) the most productive zones.
4. Transmissivities and storage coefficients estimated from pumping-test analysis are subject to many of the same errors as above, but the more carefully controlled tests should reduce their frequency and magnitude. In addition, a single test may not be representative of an entire hydrostratigraphic unit.
5. Transmissivities estimated from lithological data are usually biased to an unknown degree.

1.4 Model Construction

Ground-water models are constructed by using the types of data alluded to in the previous section. Hence, measured or estimated parameter data, either reliable or complete enough to employ directly in a model to reproduce measured head data with an acceptable model fit, are rare. As a result, adjustment of parameter values, and sometimes basic model structure, is used to improve model fit. Two basic groups of methods currently in use to accomplish this are: (1) trial and error procedures and (2) optimization methods that minimize a formal objective function.

1.4.1 Trial and Error Methods

Trial and error is the method of repeated simulation until the calculated head distribution obtained with a reasonable set of parameters fits closely enough to satisfy the analyst. Sometimes an objective measure of goodness of fit, such as $\Sigma(h^{calc}-h^{obs})^2$, is used to aid the analyst in deciding whether or not a change in parameters (or model structure) has improved the overall model fit. However, no matter how the method is applied, it has several inherent critical deficiencies:

1. No methodology exists to guarantee that the simulations will proceed in a direction that could lead to the best set of parameters.
2. Determining when that best set has been reached is difficult.
3. No practical way of determining how many other sets of parameters could yield similar correspondence between h^{calc} and h^{obs} exists.
4. Deciding whether or not additional parameters or a more refined model would significantly improve model fit is difficult.
5. No way of quantitatively assessing the predictive reliability of the model exists.

A method of model construction that addresses these deficiencies would allow construction and use of a model with a much greater degree of confidence than that provided by trial and error methods. Hence, attention is turned to formal optimization procedures.

1.4.2 Formal Optimization Procedures

Optimization procedures utilize a formal criterion of goodness of fit, often called an objective function. This function is minimized (or sometimes maximized, depending on the form of the function) with respect to the parameters to yield an optimum or best-fit solution. Minimization (or maximization) sometimes is subject to certain other criteria regarding values that the parameters, or pertinent functions of the parameters related to the model, may take on. These criteria are called constraints.

Examples of objective functions are:

$$\sum_{\ell=1}^{n_s}(h_\ell-\hat{h}_\ell)^2 \ ,$$

$$\sum_{\ell=1}^{n_s}|h_\ell-\hat{h}_\ell| \ ,$$

$$\max_{\ell}|h_\ell-\hat{h}_\ell| \ , \text{ and}$$

$$\sum_{\ell=1}^{n_s} w_\ell(h_\ell-\hat{h}_\ell)^2+\sum_{m=1}^{n_p}k_m(p_m-\hat{p}_m)^2$$

where

h_ℓ=observed head,
$\hat{h}_\ell$=calculated head,
p_m=observed or prior estimate of a parameter,
$\hat{p}_m$=calculated parameter value that, when used in the model, produces $\hat{h}_\ell$,
w_ℓ=weight related to the reliability of the observation h_ℓ,
k_m=similar weight applied to p_m,
n_s=number of observations of head, and
n_p=number of observations of parameters.

The last example is called a compound objective function because it contains both head and parameters explicitly. Note that minimization of each of the functions with respect to the parameters of the model produces a solution that is overall a best fit to the data, according to the objective function. If the signs of the

functions were changed, maximization would produce the same result.

Examples of constraints are:

$$p_m^L < \hat{p}_m < p_m^U ,$$

$$a\hat{p}_k + b\hat{p}_m + c\hat{p}_n = f ,$$

$$a\hat{p}_k + b\hat{p}_m + c\hat{p}_n < f ,$$

where a, b, c, and f are constants or known functions; superscript L refers to a lower limit; and superscript U refers to an upper limit. The best-fit solution obtained by minimizing (or maximizing) the appropriate objective function must simultaneously satisfy the appropriate constraints.

Because the solution obtained by an optimization procedure has known properties, it may be analyzed. The exact procedures used and the extent to which the model may be analyzed depend on the type of optimization method selected for use. Statistical regression procedures handle, on a probabilistic basis, the propagation of data errors (with respect to the model) into the estimates of parameters and predictive capability of the model. Methods have been developed for estimating parameters, testing assumptions made during development of techniques, testing model fit, determining the reliability and significance of the model and the parameters contained in it, effecting corrective measures for violation of some assumptions, and estimating the reliability of predictions to be made with the model. These procedures and the statistical background necessary to apply them are detailed in the remainder of the text.

References Cited

Dagan, Gedeon, 1986, Statistical theory of groundwater flow and transport—pore to laboratory, laboratory to formation, and formation to regional scale: Water Resources Research, v. 22, no. 9, p. 120S–134S.

Gelhar, L.W., 1984, Stochastic analysis of flow in heterogeneous porous media, *in* Bear, Jacob, and Corapcioglu, M.Y., eds., Fundamentals of transport phenomena in porous media: Dordrecht, The Netherlands, Martinus Nijhoff, p. 673–720.

_____1986, Stochastic subsurface hydrology from theory to applications: Water Resources Research, v. 22, no. 9, p. 135S–145S.

Neuman, S.P., 1980, A statistical approach to the inverse problem of aquifer hydrology, 3—Improved solution method and added perspective: Water Resources Research, v. 16, no. 2, p. 331–346.

Additional Reading

Freeze, R.A., and Cherry, J.A., 1979, Groundwater: Englewood Cliffs, N.J., Prentice-Hall, p. 15–77, 356–359.

2 Review of Probability and Statistics

Casual observation of our environment indicates that many phenomena are not strictly predictable. We cannot, for instance, exactly say what the maximum air temperature at any particular location will be tomorrow, although we might be able to give a probable range. This probable range might be based on our past experience, which would enable us to say that tomorrow's high, considering the location and season, will probably fall within a specified interval. A more sophisticated forecasting model may enable us to reduce the range within which we think tomorrow's high will fall, but random elements in the forecasting procedure would preclude giving an exact answer. As another example of randomness, consider the toss of a coin. Prior to the toss, we can only give the possible outcomes, either a head or a tail, and, if the coin is fair, say that either have equal likelihood of occurring. However, this ability to state precisely that any future outcome of this experiment can, with equal probability, result in either a head or a tail is an important advantage over that offered for predicting tomorrow's maximum temperature. In this latter case, because of the complex nature of the processes resulting in tomorrow's maximum, the likelihood that we could give a precise statement concerning the probability that it will fall in our predicted interval is remote. Instead of attempting to untangle these complexities, we might opt to study the history of maximum temperatures at the location and annual date in question. By assuming that this history will extrapolate into the future (that is, that weather dynamics in future years will remain essentially unchanged from those in previous years), we could give an estimate of the likelihood that tomorrow's maximum will fall in a particular interval. However, tools need to be developed to carry out this investigation.

2.1 Basic Concepts

Randomness itself can be considered to be centered around an experiment; the outcome of the experiment will have a random quality attached to it. For example, in a coin-toss experiment, the outcome is dominated by the random element (either a head or a tail). On the other hand, many experiments have a large deterministic factor. For example, in a chemical titration experiment we measure the unknown and the amount of titrant used, then calculate the amount of a specific substance in the unknown. However, measurement error creeps into our technique, and results vary from realization to realization of the experiment. Some experiments, such as annual, peak river flows, are not ours to perform but only to observe. This experiment is an example of an event in nature that has a large random component which nature provides. As we attempt to measure these flows, we introduce additional randomness, which we generally ignore. Hydraulic conductivities measured from core samples are similar to peak flows; nature has already provided for randomness, which is constrained by certain deterministic factors, such as type of source material, distance of transport, climate, and diagenesis. Again, for every realization of this experiment, measurement error is introduced, which may not be small.

All possible outcomes of an experiment are known as its sample space. The sample space of a coin-toss experiment consists of either a head (H) or tail (T):

$$S=\{H,T\}.$$

If the experiment consists of the toss of two coins, then the sample space consists of

$$S_1=\{(H,H),(H,T),(T,H),(T,T)\}.$$

On the other hand, if we are only interested in the total number of heads which might result from a single toss of two coins, we could define the experiment as this sum, which would result in the sample space

$$S_2=\{0,1,2\}.$$

In the case of S_1, every member of the sample space is equally likely to occur, whereas for S_2, a one is twice as likely to occur as either zero or two, provided that the coin is fair.

The sample space for a hydraulic-conductivity experiment could be defined as all positive real

numbers; that is, measurements from cores might result in values (outcomes) which could be as small as zero or, if we stretch our imaginations, infinitely large. This space could be considered to be a continuous equivalent of the S_2 space for the two-coin experiment. That is, a porous medium is an extremely complex random process itself. By conducting hydraulic-conductivity measurements on cores, we quantify this randomness in much the same way that counting heads quantifies an outcome of the two-coin experiment. However, by quantifying the randomness of the porous medium in this manner, we have never investigated the possible existence of more basic, perhaps nonnumeric sample spaces similar to S_1 of the two-coin experiment for a porous medium. Even if we were to discover the existence of such a space, we would then need to find a rule, or algorithm, which would allow us to connect the two spaces. We shall not worry about the possibilities of an S_1-like space for many processes; however, when they are available, they provide an excellent mechanism for investigating the characteristics of S_2-like spaces.

An event is defined as any subset of the sample space. The investigator is usually interested in the relative frequency of occurrence of an event. In the case of the S_2 space and the two-coin experiment, it is apparent that half the time a realization experiment should result in a one. This event is equivalent to the event in the S_1 space corresponding to the union of (H,T) and (T,H), which occurs with a relative frequency of one-half. Thus, the relative frequency of a head occurrence for the two-coin experiment is not dependent upon the definition of the sample space, but on the basic randomness controlling the experiment.

The investigator is frequently confronted with the problem of needing a numerical result for the outcome of a random, but not necessarily numerical, experiment. In the case of coin-toss experiments, the basic outcome is seen to be a particular arrangement of heads and (or) tails. By assigning a head a value of one and a tail a value of zero and then summing, it is possible to translate these basic results into something measurable. This process of assigning a numerical value to a nonnumerical outcome leads to the definition of a random variable.

Definition: A random variable is a function whose value is a real number determined by each element in a sample space.

When the outcome of the experiment is numerical, then this result can be considered to be the random variable (this statement is merely a special case of the above definition). From the above definition, we see that a mathematical transformation of a random variable is also a random variable. (Throughout this review, a random variable is indicated by an upper case English or Greek letter, whereas a value that it may take on is indicated by another letter, usually lower case of the same type as used for the random variable.)

The concepts of a random experiment, sample space, and random variable are flexible. For instance, if in the case of the toss of two coins, the experiment is defined as the total number of heads appearing, then the S_2 sample space is an automatic result, and the random variable can also be considered to be this result. However, if the experiment is defined to be the arrangement of heads and (or) tails resulting from a toss (that is, the S_1 space), then the same effect can be obtained by letting the random variable over the S_1 space be a function that assigns a one to a head and a zero to a tail and then sums the result. The investigator usually defines the sample space, or experiment, to suit a particular objective. As a matter of convenience, the space is usually selected such that the relative frequencies of occurrence of events within the space are definable. Access to such basic sample spaces as S_1 for the two-coin experiment allow for the calculation of relative frequencies for events in both S_1 and S_2. Without the existence of a space like S_1, determining the true relative frequency of occurrence for an event in S_2 is difficult, if not impossible. This situation is also evident from the hydraulic-conductivity experiment, where only an S_2-like sample space is available to the investigator.

A random variable can also be described as either being discrete, as in the coin-toss experiment, or continuous, as represented by the hydraulic-conductivity experiment. A discrete random variable is defined over a sample space whose elements are discrete, although there may be as many as there are whole numbers

(mathematicians refer to this phenomenon as being countably infinite). A continuous random variable is defined over a continuous sample space whose elements are infinite in number (therefore these elements are uncountably infinite).

2.2 Frequencies and Distributions

2.2.1 Discrete Random Variables

Although frequencies of occurrence are usually associated with events in a sample space, they are also associated with values of random variables, since random variables are functions of the elements in a sample space. That is, particular values of a random variable correspond to particular events in the sample space and, therefore, have frequencies of occurrence. Even though we will speak of the relative frequency of occurrence for particular values of a random variable, we are, in reality, speaking of a corresponding event in the sample space. In fact, we frequently use a range of values of a random variable to define an event in a sample space, thus avoiding the task of describing which elements of the sample space compose the event.

Frequencies of occurrence for events in many discrete sample spaces can be deduced from the following axiomatic premise: If an experiment can result in any one of N different equally likely outcomes, and if exactly n of these outcomes correspond to event A, then the relative frequency of occurrence of A is n/N. As a simple example of employment of this premise, consider an experiment consisting of a toss of a die. The sample space consists of the integers 1 through 6 and, for any realization of the experiment, each element of the sample space has equal likelihood of occurrence. By considering each element of the sample space to be an event, one can calculate the frequency of occurrence, $f(x_i)$, with which a random variable takes on the value x_i. For this experiment, only the integer values 1 through 6 of x_i have frequencies of occurrence other than 0; $f(x_i)$ can be graphically represented as shown in figure 2.2-1. In this case, $f(x_i)$ is referred to as the discrete density function of the discrete random variable

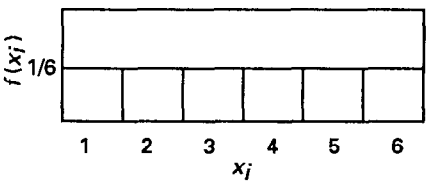

Figure 2.2-1

consisting of the outcome of a toss of a single die.

When two dice are cast, the experiment can be defined either as the sum that results from the toss or simply as all possible arrangements that could appear on the dice. If the sum is chosen, then the sample space consists of the integers 2–12, which would also be the range of values that the random variable could take on. The elements of this space, however, are not equally likely to occur. The sample space consisting of all arrangements of the numbers appearing on the two dice, presented graphically in table 2.2-1, has elements which are equally likely to occur.

Table 2.2-1

Second die	First die 1	2	3	4	5	6
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

The relative frequency of occurrence of an event corresponding to any subset of elements in this space can be calculated by using the premise concerning equally likely outcomes.

A random variable, consisting of the sum that results from any outcome of the two dice experiment, takes on the integer values 2–12 over the sample space represented by table 2.2-1. The discrete density function for this random variable can now be derived from the basic premise concerning outcomes that are equally likely, since each value for this discrete random variable corresponds to a particular event consisting of a particular subset of elements in the

sample space indicated by table 2.2–1. Thus, the value of $x_i=3$ corresponds to the event containing the elements (2,1) and (1,2) and has a relative frequence of occurrence of 2/36. Letting x_i represent the integer values that this random variable can obtain, its density function, $f(x_i)$, can be represented as shown in figure 2.2–2.

Note that had the first definition of the experiment been used, then every element of the sample space consisting of the integers 2–12 would have frequencies of occurrence, when considering each element as an event, equivalent to those shown in figure 2.2–2.

Frequencies of occurrence, or deduced frequencies of occurrence as indicated in figures 2.2–1 and 2.2–2, are indications of the future. We can make probability statements concerning the possibility of a random variable taking on future values from such knowledge. In a craps (two-dice) game, we know that the probability of rolling a natural, an outcome of 7 or 11 on the first cast, is 2/9 simply because these values of the random variable for the two-dice experiment correspond to elements in the sample space which occur with a relative frequency of 2/9. Formally, the statement that this discrete random variable X take on the values of 7 or 11 with a probability of 2/9 is written

$$P(X=7 \text{ or } X=11)=2/9.$$

The probability that this random variable takes on any integer value between 2 and 12 is obtainable directly from its frequency density, figure 2.2–2.

A probability statement that is frequently encountered concerns the probability that a random variable is less than or equal to a specific value. For the random variable corresponding to the sum of outcomes of the cast of two dice, we may ask, what is the probability that the random variable X is less than or equal to 5? The probability of this event is equal to the probability that X take on any integer value 2 through 5:

$$P(X\leq 5)=P(X=2 \text{ or } X=3 \text{ or } X=4 \text{ or } X=5).$$

This probability is the sum of the probabilities of the individual events that X take on the integer values 2 through 5:

$$P(X\leq 5)=1/36+2/36+3/36+4/36=5/18.$$

(If the student is not convinced of this relationship, he or she should examine the elements of the sample space represented by table 2.2–1 to ascertain that it holds.) Note that $P(X\leq 12)$ is unity; that is, an event which occurs with a probability of one will, undoubtedly, take place. A probability of zero indicates, on the other hand, that the event of concern cannot possibly occur.

The probability statement $P(X\leq a)$, where a is any real number, is given a special definition for both discrete and continuous random variables. That is, $F(a)=P(X\leq a)$ is known as the

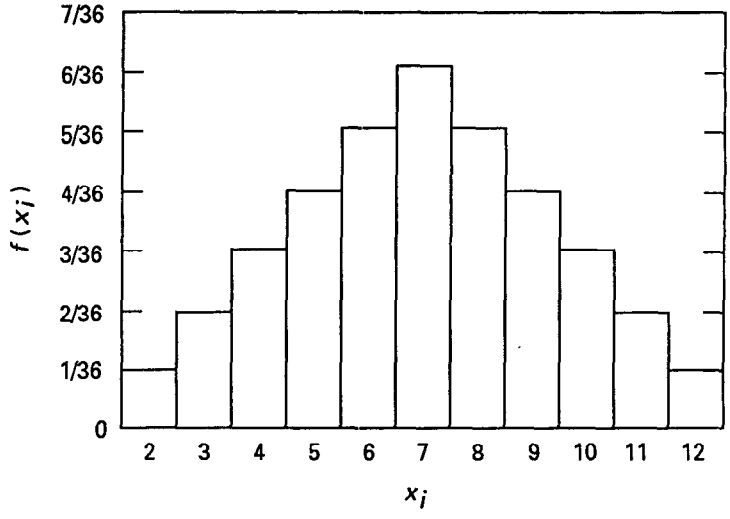

Figure 2.2–2

cumulative distribution function of the random variable X. For the case of the sum of outcomes for two dice, $F(a)$ appears as illustrated in figure 2.2-3. Because a random variable represents a functional mapping from the sample space to the real number space, we can be assured that the probability of the event $X \leq a$ exists and is equal to the sum of the probabilities of all events corresponding to values of the random variable which are less than or equal to a. In general, for discrete random variables, the cumulative distribution function can be evaluated by summing the appropriate relative frequencies of occurrence:

$$F(a) = \sum_{x_i \leq a} f(x_i). \qquad (2.2\text{-}1)$$

The cumulative distribution function for all random variables, discrete or continuous, has the following properties:

1. $F(a)$ is a nondecreasing function of a,
2. $\lim_{a \to \infty} F(a) = 1$,
3. $\lim_{a \to -\infty} F(a) = 0$.

These properties will be demonstrated in detail for continuous random variables in a later section. For a discrete random variable, these properties reflect the fact that, by definition, the discrete density function can never have a negative frequency of occurrence and that the sum of frequencies must equal one.

In the next section, an estimator for the density function of continuous random variables is developed, which will eventually allow us to explore the nature of density and cumulative distribution functions of continuous random variables.

Problem 2.2-1

An urn contains one red, one white, and two blue balls, all of equal dimensions. A ball is

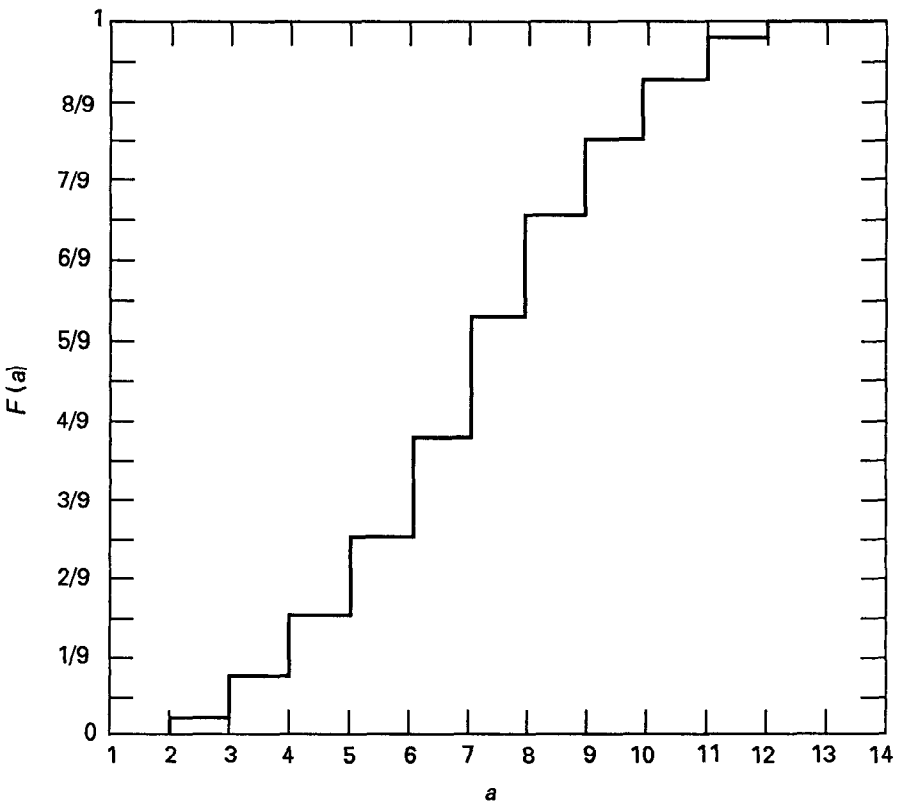

Figure 2.2-3

drawn from the urn, replaced, and then another draw is made.

a. What possible arrangements (red, white, and (or) blue) of the two balls, considering order of selection, could occur (see, for example, table 2.2-1 for two dice)?
b. What is the frequency of occurrence of any of the above events? (Hint: let the balls be represented by the symbols R, W, $B1$, and $B2$.)
c. A value of one is assigned to a blue ball, two to a red ball, and three to a white. A random variable consists of the sum of any outcome consisting of two draws with replacement. Develop a discrete density function for this random variable.
d. What is the probability that this random variable takes on a value of 4? What arrangements of balls correspond to this value of the random variable?

2.2.2 Histograms

In many cases, we do not have access to all values of random variables in a sample space (in particular, for many continuous random variables). We sample the population consisting of all possible values of the random variable and hope to draw inferences from this sample. The inferences we draw are usually in the form of statistics, which we refer to as sample statistics. We like to think that sample statistics estimate values of population parameters, which are constants reflecting the true frequency distribution of the random variable. This is frequently the case if the observations composing samples are made randomly and without bias. Samples composed of such observations are referred to as random samples and are expected to be representative of the population.

Estimates of density functions for random variables are frequently made from random samples. Although certain experiments, such as a coin toss, allow for the deduction of frequencies of occurrence of events, other experiments defy a theoretical calculation, forcing us to estimate from a random sample. These estimates, known as histograms, are generally constructed by repeating the experiment a large number of times (thus, sampling the population of all possible outcomes), dividing the range of these outcomes into class intervals, and calculating the relative number of points that fall in each interval. We might imagine, for example, that we could watch a craps game and note the outcome of each roll of dice. After a thousand rolls, we would calculate the relative percentage of each integer, 2-12, which occurred. If these sample frequencies of occurrence were not close to that shown previously for the theoretical result, we would suspect that the dice had been tampered with.

As an example of a histogram constructed from observed values of a continuous random variable, consider the transmissivity data shown in table 2.2-2. Figure 2.2-4 represents a histogram constructed directly from these data, which constitute a random sample from the population of transmissivities as determined from specific capacities of wells in carbonate rocks of central Pennsylvania. A second histogram, figure 2.2-5, was constructed from a logarithmic transformation of these data as shown in table 2.2-3. The first histogram was constructed by using a class interval of 50,000 gal/d/ft, and the second is based upon an interval of one-half a $\log_{10}$ cycle. The first histogram is not very illustrative because most of the wells have transmissivities less than 50,000 gal/d/ft (the underlying population frequency is probably heavily skewed to the right). By logarithmically transforming of the random variable, we scale the abscissa so as to remove the skewness in the histogram, causing it to be more bell shaped. This type of transformation is frequently used on random variables that have a zero lower bound, causing the transformed variable to have tails that tend to infinity in both directions. The transformation also tends to remove any right skewness in the frequency distribution of these random variables. With regard to the transformed variate, the histogram in figure 2.2-5 suggests a bell-shaped population frequency distribution. More data and smaller class intervals, as suggested in the following paragraphs, should cause the histogram shown in figure 2.2-5 to approach its population shape, which we may suspect to be a normal distribution; the untransformed random variable would then result from a log-normal distribution.

Table 2.2-2

[From Siddiqui (1969, p. 433–436)]

Transmissivity gal/d/ft	$Log_{10}T$	Transmissivity gal/d/ft	$Log_{10}T$
15.0	1.176	2,370.0	3.375
18.0	1.255	2,440.0	3.387
21.0	1.322	2,540.0	3.405
29.0	1.462	2,800.0	3.447
32.0	1.505	2,820.0	3.450
35.0	1.544	3,380.0	3.529
50.0	1.699	4,410.0	3.644
52.0	1.716	4,520.0	3.655
56.0	1.748	5,500.0	3.740
62.0	1.792	5,650.0	3.752
84.0	1.924	6,030.0	3.780
92.0	1.964	6,240.0	3.795
106.0	2.025	6,340.0	3.802
118.0	2.072	7,290.0	3.863
142.0	2.152	8,130.0	3.910
160.0	2.204	11,000.0	4.041
175.0	2.243	13,100.0	4.117
184.0	2.265	13,700.0	4.137
202.0	2.305	14,500.0	4.161
264.0	2.422	17,200.0	4.236
354.0	2.549	17,700.0	4.248
370.0	2.568	19,700.0	4.294
374.0	2.573	23,100.0	4.364
455.0	2.658	24,200.0	4.384
463.0	2.666	26,400.0	4.422
515.0	2.712	33,400.0	4.524
528.0	2.723	34,700.0	4.540
615.0	2.789	42,400.0	4.627
705.0	2.848	46,300.0	4.666
753.0	2.877	52,000.0	4.716
800.0	2.903	66,500.0	4.823
984.0	2.993	68,400.0	4.835
1,150.0	3.059	132,000.0	5.121
1,290.0	3.111	152,000.0	5.182
1,500.0	3.176	423,000.0	5.626
1,580.0	3.199	423,000.0	5.626
1,670.0	3.223	528,000.0	5.723
1,850.0	3.267	528,000.0	5.723
2,310.0	3.364	528,000.0	5.723

Table 2.2-3

Class interval[1]	Number of occurrences	Relative frequency	Cumulative frequency
1.0–1.5	4	0.051	0.051
1.5–2.0	8	.103	.154
2.0–2.5	8	.103	.257
2.5–3.0	12	.154	.411
3.0–3.5	12	.154	.565
3.5–4.0	10	.128	.693
4.0–4.5	10	.128	.821
4.5–5.0	7	.090	.911
5.0–5.5	2	.026	.937
5.5–6.0	5	.064	1.001
Total	78		

[1]Based on $log_{10}T$, table 2.2-2

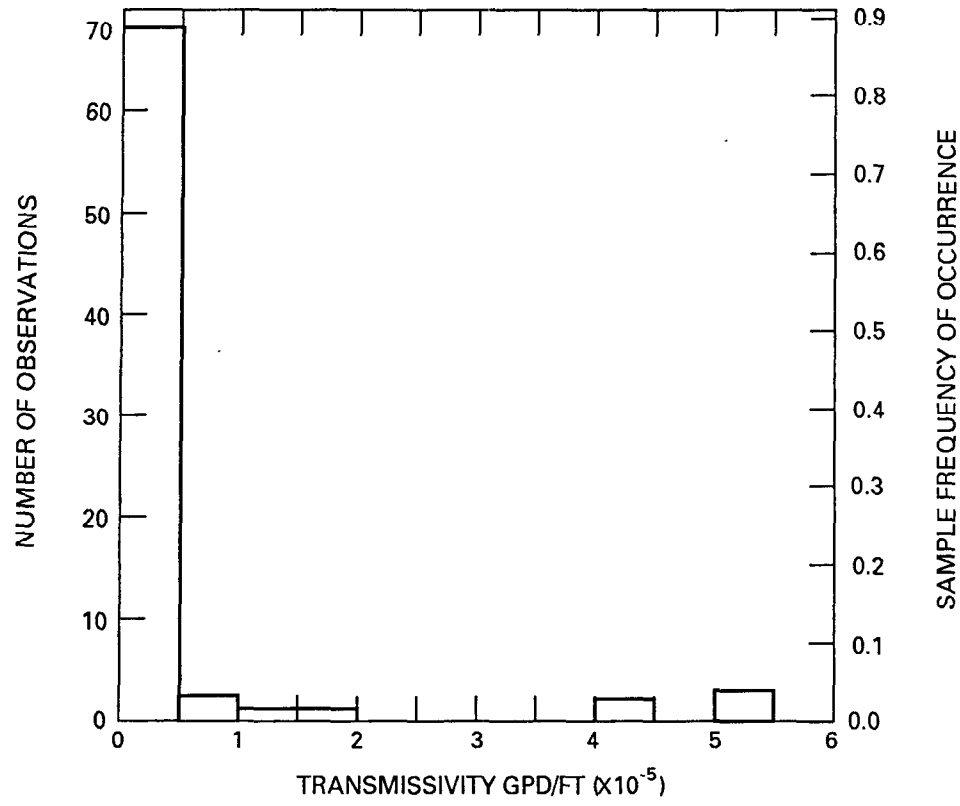

Figure 2.2–4

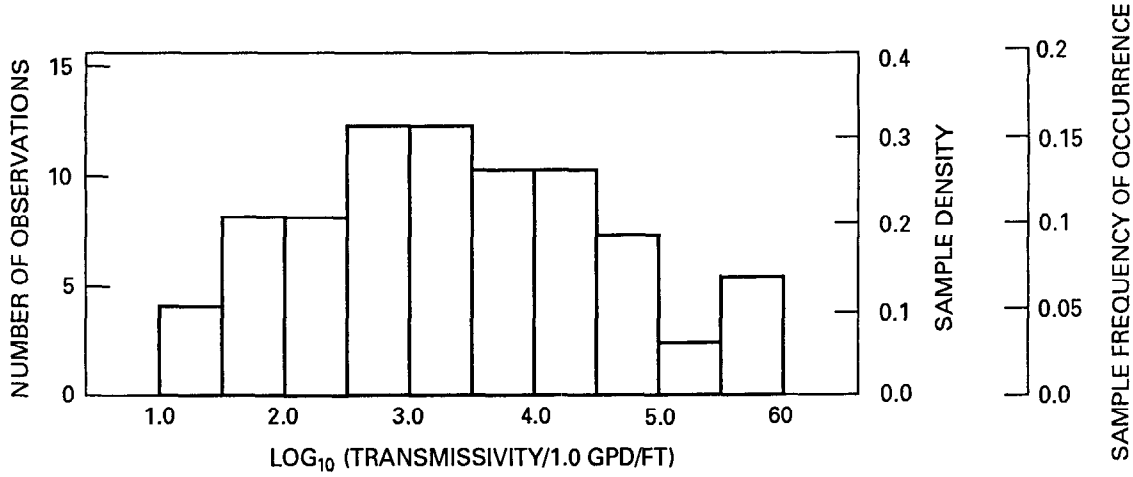

Figure 2.2–5

We are now in position to estimate the probability of occurrence of an event associated with the log-transformed random variable. Let $X=\log_{10}T$ represent the transformed variate plotted in figure 2.2-5, and assume that the histogram for X is representative of the population frequency. The estimated probability that X is less than or equal to 5.0 but greater than 4.5, $P(4.5<X\leq 5.0)$, then, is the sample frequency of occurrence of this event (equal to 0.09). The probability that X is less than or equal to 5.0 can be estimated by summing the frequencies of occurrence of all events smaller than 5.0; thus $P(X\leq 5.0)\approx 0.911$. Thus, the chances are about 91 in 100 that the transmissivity of the carbonate rocks in central Pennsylvania, as determined by any random well, will be less than or equal to 1×10^5 gal/d/ft. The reader should realize that these results are only approximate, as the histogram is an approximation of the true population frequency distribution.

An estimate of the cumulative distribution function can also be constructed from a random sample. Let $F_n(a)$ represent this estimate, known as the sample distribution function; an appropriate estimator for $F_n(a)$ is the sum of all estimated relative frequencies for values of the random variable X less than a:

$$P(X\leq a)\approx F_n(a)=\sum_{i\leq a/\Delta x} f_i^* \qquad (2.2\text{-}2)$$

where

$f_i^*=n_i/n$=sample frequency of occurrence of an event represented by the ith class interval,
Δx=size of class interval,
n_i=number of outcomes having values in interval i, and
n=size of random sample.

An application of this procedure for the logarithmic transformation of transmissivity is shown in figure 2.2-6.

2.2.3 Continuous Random Variables

The definition of frequency f_i^* used in equation 2.2-2 suffers from the deficit that it is dependent upon the size of the class interval; that is, if Δx decreases in size while n remains constant, then f_i^* must also decrease, as we are also effectively decreasing the value of n_i within this interval. Indeed, even if n were allowed to become large as Δx decreases, thus causing n_i for any arbitrary interval to be large, f_i^* could still be made arbitrarily small by decreasing the interval size sufficiently. However, this phenomenon would prevent us from defining a frequency for a single point in a continuous

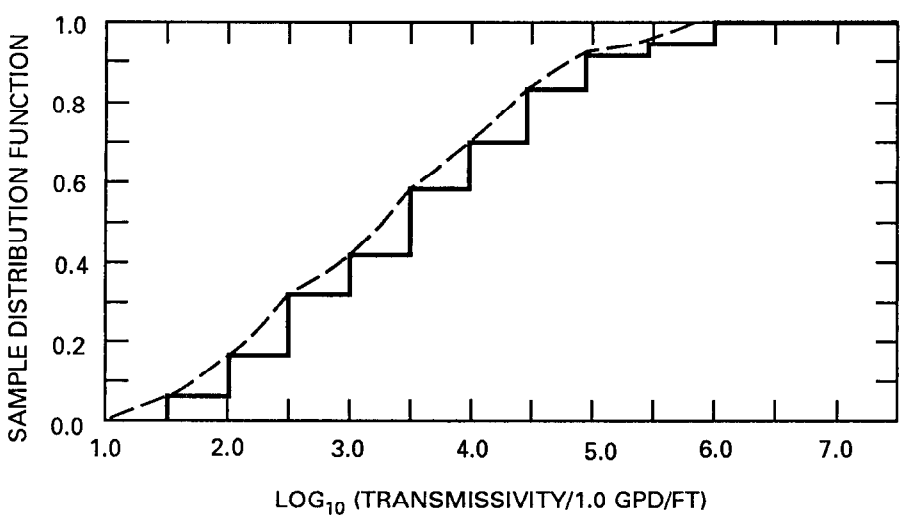

Figure 2.2-6

random variable, unless we are content to associate it with some arbitrary class interval. To overcome this problem, probabilists have defined a different measure of frequency for continuous random variables that consists of the frequency of occurrence f_i^* scaled by its class interval:

$$f_i = f_i^*/\Delta x \quad . \tag{2.2-3}$$

This normalized frequency, referred to as the sample density, should be relatively stable for reasonable choices of Δx and n, and in the limiting case of n approaching infinity and Δx approaching zero, f_i should be constant. An additional ordinate has been added to figure 2.2-5 to show the sample-density distribution of the $\log_{10}T$ data.

The sample distribution function of equation 2.2-2 can now be redefined in terms of equation 2.2-3 as follows:

$$F_n(a) = \sum_{i \le a/\Delta x} f_i \Delta x \quad . \tag{2.2-4}$$

This definition lends itself to an exploration of the population equivalents of $F_n(a)$ and f_i. If the random sample is of sufficient size to sample every member of the sample space and Δx is taken infinitely small, then the population equivalents of $F_n(a)$ and f_i should be approached. By letting n become large and Δx small, we see that

$$\sum_{i \le a/\Delta x} f_i \Delta x \approx \int_{-\infty}^{a} f(x)dx \tag{2.2-5}$$

where $f(x)$, the population equivalent of f_i, is known as the probability density function. Because $f(x)$ is the population equivalent of f_i, then the integral representation in equation 2.2-5 of summing these scaled frequencies must be the population equivalent of $F_n(a)$, which of course is the same cumulative distribution function defined earlier in section 2.2.1:

$$F(a) = \int_{-\infty}^{a} f(x)dx = P(X \le a) \quad . \tag{2.2-6}$$

However, because a random sample, whether it be finite or infinite, is countable, equation 2.2-5 must be given a special interpretation. Note that, because $f(x)$ is the continuous analog of f_i, it is always a non-negative quantity.

A stronger statement than equation 2.2-5 can be made concerning the equivalence of $F(a)$ and $F_n(a)$ for large sample sizes by noting that $F_n(a)$, prior to sampling, is a random variable. That is, if we were to collect different samples of the same size n from the same population, we would not expect that $F_n(a)$, computed from each random sampling, would have the same value. We would only hope that, as n becomes large, these different values would approach some constant. Indeed, probabilists have shown that, with a probability of one, $F_n(a)$ becomes the constant $F(a)$ as n goes to infinity. This result is particularly remarkable if we first consider that $F_n(a)$ can only take on a countable number of values k/n, $0 \le k \le n$, where k is an integer (see equation 2.2-2). Thus, although the values of $F(a)$ are uncountably infinite (continuous), $F_n(a)$ can only be, in the case that the random sample is infinitely large, at most, countably infinite. We will use this result loosely by allowing equation 2.2-5 to take on the indicated limits,

$$\lim_{\substack{\Delta x \to 0 \\ n \to \infty}} F_n(a) = F(a) \tag{2.2-7}$$

and noting that this result only can occur with a probability of one.

Both $f(x)$ and $F(x)$ are continuous functions of values of the random variable X. For the previously illustrated case of $X = \log_{10}T$, the density function might appear as in figure 2.2-7. Figure 2.2-7 represents the population equivalent of figure 2.2-5, as if all possible outcomes of the random variable were available to us. Similarly, the cumulative frequency distribution, the population equivalent of figure 2.2-6, for this random variable might appear as in figure 2.2-8.

Because of equation 2.2-6, the density function $f(x)$ can be defined in terms of the cumulative distribution function $F(x)$ by differentiation:

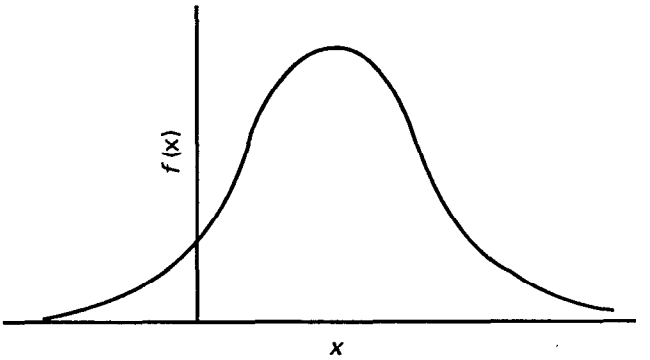

Figure 2.2-7

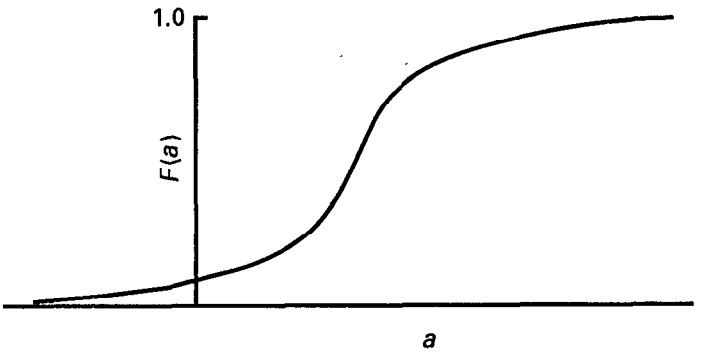

Figure 2.2-8

$$f(a)=\frac{dF(a)}{da}=\frac{d}{da}\int_{-\infty}^{a} f(x)dx. \qquad (2.2\text{-}8)$$

This result follows directly from the fundamental theorem of integral calculus, and is applicable only to density functions of continuous random variables. Equation 2.2-8 is one of three concepts which defines density functions of continuous random variables. The other two state that $f(x)$ must be greater than or equal to zero for any possible value of the random variable and, as will be demonstrated in the next section, that the total mass under the frequency curve must be unity. All density functions of continuous random variables have these concepts in common.

Problem 2.2-2

a. Construct histograms for the following specific-conductance data using class intervals of 100 and 200 μmho/cm, such that the abscissa and ordinate of both histograms are scaled equally. What is the effect of changing the class interval?

b. Construct a cumulative frequency distribution from your 100 μmho/cm class-interval results. Let X represent the specific-conductance random variable; what is

$P(X\leq 600)$?
$P(X>400)$?
$P(400<X\leq 600)$?
$P(X\leq 1300)$?

Ordered specific-conductance data

[Data in μmho/cm for wells in carbonate rocks of Maryland. From Nutter, 1973, p. 63–68]

63	423	501	582	685	836
76	433	504	596	697	839
168	439	509	598	700	876
278	440	512	600	704	882
301	440	518	604	710	895
304	440	518	617	721	897
310	444	527	620	723	904
315	452	529	627	724	906
319	452	533	629	726	915
323	452	537	632	728	948
332	456	538	636	740	968
347	462	542	641	750	969
357	469	552	647	750	982
359	471	562	659	764	997
363	473	564	659	765	1,030
389	477	564	661	779	1,080
407	487	565	664	783	1,106
408	490	566	665	789	1,120
411	492	570	670	808	1,170
413	493	575	673	808	1,230
417	493	578	675	815	
418	499	582	677	820	

2.2.4 Properties of Cumulative Distribution Functions

In the previous section, the cumulative distribution function $F(a)$, defined by the probability statement $P(X \leq a)$, was noted to have the integral form of equation 2.2–6 for continuous random variables. We state all manner of probability statements in terms of the cumulative distribution function, as this is a standard form. For this purpose, properties of cumulative distribution functions, with applications to other probability statements, are developed in this section.

The probability that a random variable X takes on a value in the interval $(a, b]$ can be expressed in terms of cumulative distribution functions as

$$P(a<X\leq b)=\int_a^b f(x)dx=F(b)-F(a) \quad . \qquad (2.2\text{–}9)$$

This statement is a direct result of integral calculus, whereby integration is used to sum all the frequencies of occurrences of values of the random variable between a and b. From equation 2.2–9 one sees that the cumulative distribution function is a nondecreasing function of x, because

$$0\leq P(a<X\leq b)\leq 1.$$

That the total mass under the sample density curve f_i is unity is evident from equation 2.2–4; that is,

$$\lim_{a\to\infty} F_n(a)=\lim_{a\to\infty} \sum_{i\leq a/\Delta x} f_i\Delta x=1 \quad . \qquad (2.2\text{–}10)$$

Because the probability density function $f(x)$ of a continuous random variable X is a limiting form of the sample density f_i, the mass under its curve is also unity:

$$\lim_{a\to\infty} F(a)=\lim_{a\to\infty} \int_{-\infty}^{a} f(x)dx=1 \quad . \qquad (2.2\text{–}11)$$

Equation 2.2–11 is a property of all cumulative distribution functions. Similarly,

$$\lim_{a\to-\infty} F(a)=\lim_{a\to-\infty} \int_{-\infty}^{a} f(x)dx=0 \quad , \qquad (2.2\text{–}12)$$

which follows from integral calculus, is also a property of cumulative distribution functions.

Equation 2.2-11 allows one to express $P(X>a)$ as

$$P(X>a)=\int_a^{\infty} f(x)dx=1-\int_{-\infty}^{a} f(x)dx=1-F(a), \quad (2.2\text{-}13)$$

which is also a result of Riemannian integration. An alternate statement of equation 2.2-13 is that $P(X>a)=1-P(X\leq a)$.

By considering equation 2.2-9 in a limit form, we can also find the probability that $X=a$:

$$P(X=a)=\lim_{\Delta x\to 0} P(a<X\leq a+\Delta x) \ ,$$

$$=\lim_{\Delta x\to 0} [F(a+\Delta x)-F(a)]=0. \quad (2.2\text{-}14)$$

This result is unique to continuous random variables, in contradistinction to discrete random variables. From equation 2.2-14 one sees that $P(X\leq a)$ is equivalent to $P(X<a)$ for continuous random variables, as the endpoint, a, of the semi-infinite interval does not contribute mass to the probability statement.

Equations 2.2-9, 2.2-11, and 2.2-13 can be demonstrated for discrete random variables by using the summation form of the cumulative distribution function (equation 2.2-1). In contradistinction to continuous random variables, the endpoint in $P(X\leq a)$ for a discrete random variable can contribute significant mass to the statement.

A number of frequency densities that result from randomness in nature, or probabilistic models of random events, have been investigated and published. Cumulative distributions of these densities are frequently tabulated and are found in many reference books on probability and statistics. Equations 2.2-9 and 2.2-13 are especially useful in evaluating probability statements of tabulated random variables.

2.2.5 An Example: The Normal Distribution

Let the random variable Y represent the amount of titrant used in a titration experiment to neutralize measured amounts of the unknown x. A scatter diagram of titrant versus unknown might appear as in figure 2.2-9. The solid line represents the true stoichiometric balance between titrant and unknown. The dots, representing repetitions of the experiment, deviate from this line by an amount ϵ, which represents a value of the measurement error $\mathcal{E}$. These errors represent a continuous random variable that could theoretically vary from $-\infty$ to $+\infty$ (the graphed points only represent a random sample from the population). If the experimental apparatus is functioning properly, however, we would expect these dots to be concentrated in the general vicinity of the solid line.

A distribution that is frequently used to model errors that are symmetrically distributed about some common point is the normal

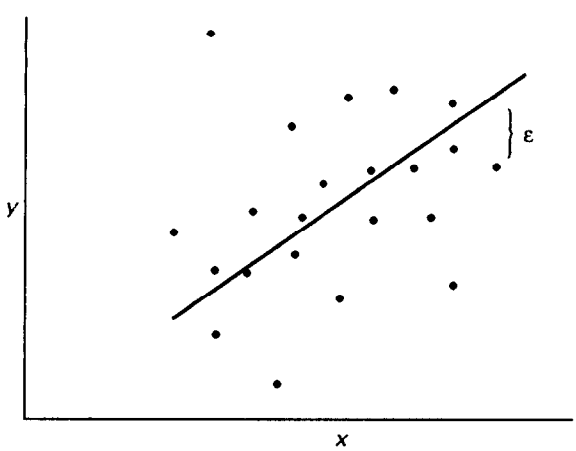

Figure 2.2-9

distribution. The density of the normal distribution is a bell-shaped curve, symmetric about its mean μ_ε, and with most of the mass concentrated within one standard deviation σ_ε of the mean (see figure 2.2-10). In the case of the titration experiment, we would hope that the most frequently found value of the error would be near-zero and expect that μ_ε would equal zero. The standard deviation σ_ε is a measure of the dispersion, or spread, of the errors about the mean and is equal to the distance from the mean to an inflection point on the curve $f(\epsilon)$. The mean and standard deviation will be formally defined in a later section.

A normal random variable is frequently standardized with its mean and standard deviation by the following transformation:

$$Z=(\mathcal{E}-\mu_\varepsilon)/\sigma_\varepsilon \quad . \tag{2.2-15}$$

The cumulative distribution for this standard normal random variable is tabulated (table 2.10-1) for use by the investigator, since its probability density function, $f_Z(z)$, is parameter free:

$$f_Z(z)=\frac{e^{-z^2/2}}{\sqrt{2\pi}} \quad . \tag{2.2-16}$$

Given the density function for the standard normal random variable, it is natural to inquire about the form of density, $f_\varepsilon(\epsilon)$, of the unnormalized random variable $\mathcal{E}$. Consider the cumulative frequency distribution for Z. By making the change of variables $z=(s-\mu_\varepsilon)/\sigma_\varepsilon$,

$$F_Z(a)=\int_{-\infty}^{a}\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz$$

$$=\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\int_{-\infty}^{\epsilon}\exp\left[-\left(\frac{s-\mu_\varepsilon}{\sigma_\varepsilon}\right)^2/2\right]ds \tag{2.2-17}$$

results where $\epsilon=a\sigma_\varepsilon+\mu_\varepsilon$ is a value of the unnormalized random variable. Since differentiation is the inverse operator of integration, equation 2.2-17 is differentiated with respect to ϵ to find $f_\varepsilon(\epsilon)$ (see also equation 2.2-8):

$$f_\varepsilon(\epsilon)=\frac{d}{d\epsilon}F_Z(a(\epsilon))$$

$$=\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\exp\left[-\left(\frac{\epsilon-\mu_\varepsilon}{\sigma_\varepsilon}\right)^2/2\right]. \tag{2.2-18}$$

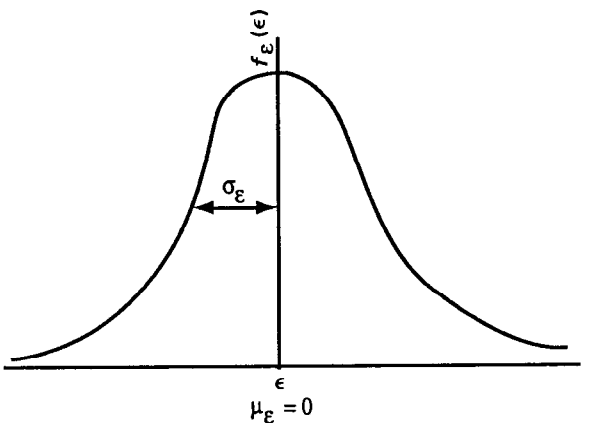

Figure 2.2-10

Note that equation 2.2-18 is not parameter free, as this density is a function of the parameters μ_ε and σ_ε.

2.3 Expectation and the Continuous Random Variable

The discussion in this section is largely presented with continuous random variables in mind. All the results, however, are applicable to discrete random variables; whenever a quantity is defined by an integration over a probability density function for the continuous case, this same quantity can almost invariably be defined by a summation over the discrete density function for the discrete case. The reader should demonstrate the veracity of this statement.

2.3.1 The Mean

The mean is a measure of central tendency of a population. As an estimator of this central tendency, consider a finite random sample consisting of n values x_i of the random variable X. If the sample frequency of occurrence f_i^* is estimated from this random sample, then a logical estimator of the central tendency is to sum the product of the central value $\bar{x}_i$ of each class interval and the frequency of occurrence for that interval:

$$\bar{x}=\sum_{i\le x_m/\Delta x} f_i^*\,\bar{x}_i=\sum_{i\le x_m/\Delta x} f_i\,\bar{x}_i\,\Delta x \tag{2.3-1}$$