Section 3

Mathematics*

Bruce A. Finlayson, Ph.D., *Rehnberg Professor and Chair, Department of Chemical Engineering, University of Washington: Member, National Academy of Engineering (Numerical methods and all general material; section editor)*

James F. Davis, Ph.D., *Professor of Chemical Engineering, Ohio State University (Intelligent Systems)*

Arthur W. Westerberg, Ph.D., *Swearingen University Professor of Chemical Engineering, Carnegie Mellon University: Member, National Academy of Engineering (Optimization)*

Yoshiyuki Yamashita, Ph.D., *Associate Professor of Chemical Engineering, Tohoku University, Sendai, Japan (Intelligent Systems)*

* The contributions of William F. Ames (retired), Georgia Institute of Technology; Arthur E. Hoerl (deceased), University of Delaware; and M. Zuhair Nashed, University of Delaware, to material that was used from the sixth edition is gratefully acknowledged.

General References: The list of references for this section is selected to provide a broad perspective on classical and modern mathematical methods that are useful in chemical engineering. The references supplement and extend the treatment given in this section. Also included are selected references to important areas of mathematics that are not covered in the *Handbook* but that may be useful for certain areas of chemical engineering, e.g., additional topics in numerical analysis and software, optimal control and system theory, linear operators, and functional-analysis methods. Readers interested in brief summaries of theory, together with many detailed examples and solved problems on various topics of college mathematics and mathematical methods for engineers, are referred to the Schaum's Outline Series in Mathematics, published by the McGraw-Hill Book Comapny.

1. Abramowitz, M., and I. A. Stegun. *Handbook of Mathematical Functions,* National Bureau of Standards, Washington, D.C. (1964).
2. Action, F. S. *Numerical Methods That Work,* Math. Assoc. of Am. (1990).
3. Adey, R. A., and C. A. Brebbia. *Basic Computational Techniques for Engineers,* Wiley, New York (1983).
4. Akai, T. *Applied Numerical Methods for Engineers,* Wiley, New York (1994).
5. Akin, J. E. *Finite Element Anaysis for Undergraduates,* Academic, New York (1986).
6. Alder, H., N. Karmarker, M. Resende, and G. Veigo. *Mathematical Programming* **44** (1989): 297–335.
7. American Institute of Chemical Engineers. "Advanced Simulators Migrate to PCs," *Chem. Eng. Prog.* **90** (Oct. 1994): 13–14.
8. American Institute of Chemical Engineers. "CEP Software Directory," *Chem. Engn. Prog.* (Dec. 1994).
9. Ames, W. F. *Nonlinear Partial Differential Equations in Engineering,* Academic, New York (1965).
10. ———. *Nonlinear Ordinary Differential Equations in Transport Processes,* Academic, New York (1968).
11. ———. *Numerical Methods for Partial Differential Equations,* 2d ed., Academic, New York (1977).
12. ———. *Ind. Eng. Chem. Fund.* **8** (1969): 522–536.
13. Amundson, N. R. *Mathematical Methods in Chemical Engineering,* Prentice Hall, Englewood Cliffs, NJ (1966).
14. Anderson, E. et al. *LAPACK Users' Guide,* SIAM (1992).
15. Antsaklis, P. J., and K. M. Passino (eds.). *An Introduction to Intelligent and Autonomous Control,* Kluwer Academic Publishers (1993).
16. Aris, R. *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysis,* vols. 1 and 2, Oxford University Press, Oxford (1975).
17. ———. *Mathematical Modelling Techniques,* Pitman, London (1978).
18. ———. *Vectors, Tensors, and the Basic Equations of Fluid Mechanics,* Prentice Hall (1962).
19. ——— and N. Amundson. *Mathematical Methods in Chemical Engineering,* vols. 1 and 2, Prentice Hall, Englewood Cliffs, NJ (1973).
20. Arya, J. C., and R. W. Lardner. *Algebra and Trigonometry with Applications,* Prentice Hall, Englewood Cliffs, NJ (1983).
21. Ascher, U., J. Christiansen, and R. D. Russell. *Math. Comp.* **33** (1979): 659–679.
22. Atkinson, K. E. *An Introduction to Numerical Analysis,* Wiley, New York (1978).
23. Atkinson, K. E. *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind,* SIAM, Philadelphia (1976).
24. Badiru, A. B. *Expert Systems Applications in Engineering and Manufacturing,* Prentice Hall, Englewood Cliffs, NJ (1992).
25. Baird, D. C. *Experimentation: An Introduction to Measurement Theory and Experiment Design,* 3d ed., Prentice Hall, Engelwood Cliffs, NJ (1995).
26. Baker, C. T. H. *The Numerical Treatment of Integral Equations,* Oxford University Press, New York (1977).
27. Barker, V. A. (ed.). *Sparse Matrix Techniques—Copenhagen 1976,* Lecture Notes in Mathematics 572, Springer-Verlag, New York (1977).
28. Beckenbach, E. F., and R. E. Bellman. *Inequalities,* 3d printing, Springer-Verlag, Berlin (1971).
29. Becker, E. B., G. F. Carey, and J. T. Oden: *Finite Elements: An Introduction,* Prentice Hall, Englewood Cliffs, NJ (1981).
30. Bellman, R. E., and K. L. Cooke. *Differential-Difference Equations,* Academic, New York (1972).
31. Bender, E. A. *An Introduction to Mathematical Modeling,* Wiley, New York (1978).
32. Bender, C. M., and Orszag, S. A., *Advanced Mathematical Methods for Scientists and Engineers,* McGraw-Hill (1978).
33. Ben-Israel, A., and T. N. E. Greville. *Generalized Inverses: Theory and Applications,* Wiley-Interscience, New York (1974).
34. Boas, R. P. Jr. *Am. Math. Mon.* **84** (1977): 237–258.
35. Bodewig, E. *Matrix Calculus,* 2d ed., Interscience, New York (1959).
36. Bogacki, M. B., Alejski, K., and Szymanowski, J. *Comp. Chem. Eng.* **13** (1989): 1081–1085.
37. Book, D. L. *Finite-Difference Techniques for Vectorized Fluid Dynamics Calculations,* Springer-Verlag, New York (1981).
38. Boor, C. de. *A Practical Guide to Splines,* Springer-Verlag, New York (1978).
39. Botha, J. F., and G. F. Pinder. *Fundamental Concepts in the Numerical Solution of Differential Equations,* Wiley, New York (1983).
40. Box, G. E. P., Hunter, W. G., and Hunter, J. S. *Statistics for Experimenters,* Wiley, New York (1978).
41. Boyce, W. E., and R. C. Di Prima. *Elementary Differential Equations and Boundary Value Problems,* 5th ed., Wiley, New York (1992).
42. Bradley, S. P., A. C. Hax, and T. L. Magnante. *Applied Mathematical Programming,* Addison-Wesley, Reading, MA (1977).
43. Brand, L. *Differential and Difference Equations,* Wiley, New York (1966).
44. Braun, M. *Differential Equations and Their Applications: An Introduction to Applied Mathematics,* 4th ed., Springer-Verlag, New York (1993).
45. Brebbia, C. A., and J. Dominguez, *Boundary Elements—An Introductory Course,* Computational Mechanics Publications, Southhampton (1988).
46. Brent, R. *Algorithms for Minimization without Derivatives,* Prentice Hall, Englewood Cliffs, NJ (1973).
47. Brigham, E. *The Fast Fourier Transform and its Application,* Prentice Hall, Englewood Cliffs, NJ (1988).
48. Bronshtein, I. N., and K. A. Semendyayev (K. A. Hirsch, trans.). *Handbook of Mathematics,* Van Nostrand (1985).
49. Brown, David C., and B. Chandrasekaran. *Design Problem Solving: Knowledge Structures and Control Strategies,* Pitman, London; and Morgan Kaufman, San Mateo, CA (1989).
50. Broyden, C. G. *J. Inst. Math. Applic.* **6** (1970): 76.
51. Brujin, N. G. de. *Asymptotic Methods in Analysis,* Dover, New York (1981).
52. Bryson, A. E., and Y-C Ho. *Applied Optimal Control,* Hemisphere Publishing, Washington, DC (1975).
53. Buck, R. C. *Advanced Calculus,* 3d ed., McGraw-Hill, New York, 1978.
54. Bulsari, A. B. (ed.). *Neural Networks for Chemical Engineers,* Elsevier Science Publishers, Amsterdam (1995).
55. Bunch, J. R., and D. J. Rose (ed.). *Sparse Matrix Computations,* Academic, New York (1976).
56. Burden, R. L., J. D. Faires, and A. C. Reynolds. *Numerical Analysis,* 5th ed., Prindle, Weber & Schmidt, Boston (1993).
57. Byrd, P., and M. Friedman. *Handbook of Elliptic Integrals for Scientists and Engineers,* 2d ed., Springer-Verlag, New York (1971).
58. Byrne, G. A., and P. R. Ponzi. *Comp. Chem. Eng.* **12** (1988): 377–382.
59. Carnahan, B., H. Luther, and J. Wilkes. *Applied Numerical Methods,* Wiley, New York (1969).
60. Carnahan, B., and J. O. Wilkes. "Numerical Solution of Differential Equations—An Overview" in *Foundations of Computer-Aided Chemical Process Design,* AIChE, New York (1981).
61. Carrier, G., and C. Pearson. *Partial Differential Equations: Theory and Technique,* 2d ed., Academic, New York (1988).
62. Carrier, G. F., and C. E. Pearson. *Ordinary Differential Equations,* SIAM (1991).
63. Carslaw, H. S. *The Theory of Fourier Series and Integrals,* 3d ed., Dover, New York (1930).
64. ——— and J. Jaeger. *Operational Methods in Applied Mathematics,* 2d ed., Clarendon Press, Oxford (1948).
65. Chamberlain, R. M., C. Lemarechal, H. C. Pedersen, and M. J. D. Powell. "The Watchdog Technique for Forcing Convergence in Algorithms for Constrained Optimization," *Math. Prog. Study* **16** (1982).
66. Chan, T. F. C., and H. B. Keller. *SIAM J. Sci. Stat. Comput.* **3** (1982): 173–194.
67. Chang, M. W., and B. A. Finlayson. *Int. J. Num. Methods Eng.* **15** (1980): 935–942.
68. Char, B. W., K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan, and S. M. Watt. *Maple V. Language Reference Manual,* Springer-Verlag, Berlin (1991).

69. Chatterjee, S., and B. Price. *Regression Analysis by Example,* 2d ed., Wiley, New York (1991).
70. Cheney, E. W., and D. Kincaid. *Numerical Mathematics and Computing,* Brooks/Cole, Monterey, CA (1980).
71. Churchill, R. V. *Operational Mathematics,* 3d ed., McGraw-Hill, New York (1972).
72. ——— and J. W. Brown. *Fourier Series and Boundary Value Problems,* 4th ed., McGraw-Hill, New York (1987).
73. ———, J. W. Brown, and R. V. Verhey. *Complex Variables and Applications,* 4th ed., McGraw-Hill, New York (1984).
74. Clarke, F. H. *Optimization and Nonsmooth Analysis,* Wiley, New York (1983).
75. Cochran, J. A. *The Analysis of Linear Integral Equations,* McGraw-Hill, New York (1972).
76. Collatz, L. *The Numerical Treatment of Differential Equations,* 3d ed., Springer-Verlag, Berlin and New York (1960).
77. Conte, S. D., and C. de Boor. *Elementary Numerical Analysis: An Algorithmic Approach,* 3d ed., McGraw-Hill, New York (1980).
78. Cooper, L., and D. Steinberg. *Methods and Applications of Linear Programming,* Saunders, Philadelphia (1974).
79. Courant, R., and D. Hilbert. *Methods of Mathematical Physics,* Interscience, New York (1953, 1962).
80. Crandall, S. *Engineering Analysis,* McGraw-Hill, New York (1956).
81. Creese, T. M., and R. M. Haralick. *Differential Equations for Engineers,* McGraw-Hill, New York (1978).
82. Cropley, J. B. "Heuristic Approach to Complex Kinetics," pp. 292–302 in *Chemical Reaction Engineering—Houston,* ACS Symposium Series 65, American Chemical Society, Washington, DC (1978).
83. Cuvelier, C., A. Segal, and A. A. van Steenhoven. *Finite Element Methods and Navier-Stokes Equations,* Reidel, Dordrecht (1986).
84. Davidon, W. C. "Variable Metric Methods for Minimization," AEC R&D Report ANL-5990, rev. (1959).
85. Davis, M. E. *Numerical Methods and Modeling for Chemical Engineers,* Wiley, New York (1984).
86. Davis, P. J. *Interpolation and Approximation,* Dover, New York (1980).
87. ——— and P. Rabinowitz. *Methods of Numerical Integration,* 2d ed., Academic, New York (1984).
88. Denn, M. M. *Stability of Reaction and Transport Processes,* Prentice Hall, Englewood Cliffs, NJ (1974).
89. Dennis, J. E., and J. J. More. *SIAM Review* **21** (1977): 443.
90. Dimian, A. *Chem. Eng. Prog.* **90** (Sept. 1994): 58–66.
91. Doherty, M. F., and J. M. Ottino. *Chem. Eng. Sci.* **43** (1988): 139–183.
92. Dongarra, J. J., J. R. Bunch, C. B. Moler, and G. W. Stewart. *LINPACK Users Guide,* Society for Industrial and Applied Mathematics, Philadelphia (1979).
93. Draper, N. R., and H. Smith. *Applied Regression Analysis,* 2d ed., Wiley, New York (1981).
94. Dubois, D., H. Prade, and R. R. Yager (eds.) *Readings in Fuzzy Sets for Intelligent Systems,* Morgan Kaufmann (1993).
95. Duff, I. S. (ed.). *Sparse Matrices and Their Uses,* Academic, New York (1981).
96. Duff, I. S. *Direct Methods for Sparse Matrices,* Oxford, Charendon Press (1986).
97. Duffy, D. G. *Transform Methods for Solving Partial Differential Equations,* CRC Press (1994).
98. Dym, C. L., and E. S. Ivey. *Principles of Mathematical Modeling,* Academic, New York (1980).
99. Edgar, T. F., and D. M. Himmelblau. *Optimization of Chemical Processes,* McGraw-Hill (1988).
100. Eisenstat, S. C. *SIAM J. Sci. Stat. Comp.* **2** (1981): 1–4.
101. Eisenstat, S. C., M. H. Schultz, and A. H. Sherman. *SIAM J. Sci. Stat. Comput.* **2** (1981): 225–237.
102. Elich, J., and C. J. Elich. *College Alegebra with Calculator Applications,* Addison-Wesley, Boston (1982).
103. Ferguson, N. B., and B. A. Finlayson. *A. I. Ch. E. J.* **20** (1974): 539–550.
104. Finlayson, B. A. *The Method of Weighted Residuals and Variational Principles,* Academic, New York (1972).
105. Finlayson, B., L. T. Biegler, I. E. Grossmann, and A. W. Westerberg. "Mathematics in Chemical Engineering," *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. B1, VCH, Weinheim (1990).
106. Finlayson, B. A. *Nonlinear Analysis in Chemical Engineering,* McGraw-Hill, New York (1980).
107. Finlayson, B. A. *Numerical Methods for Problems with Moving Fronts,* Ravenna Park Publishing, Seattle (1992).
108. Fisher, R. C., and A. D. Ziebur. *Integrated Algebra, Trigonometry, and Analytic Geometry,* 4th ed., Prentice Hall, Englewood Cliffs, NJ (1982).
109. Fletcher, R. *Computer J.* **13** (1970): 317.
110. Fletcher, R. *Practical Methods of Optimization,* Wiley, New York (1987).
111. Forsythe, G. E., M. A. Malcolm, and C. B. Moler. *Computer Methods for Mathematical Computations,* Prentice Hall, Englewood Cliffs, NJ (1977).
112. Forsyth, G., and C. B. Moler. *Computer Solution of Linear Algbraic Systems,* Prentice Hall, Englewood Cliffs (1967).
113. Fourer, R., D. M. Gay, and B. W. Kerninghan. *Management Science* **36** (1990): 519–554.
114. Friedman, N. A. *Calculus and Mathematical Models,* Prindle, Weber & Schmidt, Boston (1979).
115. Gantmacher, F. R. *Applications of the Theory of Matrices,* Interscience, New York (1959).
116. Garbow, B. S., J. M. Boyle, J. J. Dongarra, and C. B. Moler: *Matrix Eigensystem Routines—EISPACK Guide Extensions,* Springer-Verlag, Berlin and New York (1977).
117. Gear, G. W. *Numerical Initial Value Problems in Ordinary Differential Equations,* Prentice Hall, Englewood Cliffs, NJ (1971).
118. Gellert, W., H. Küstner, M. Hellwich, H. Kästner (ed.). *The VNR Concise Encyclopedia of Mathematics,* Van Nostrand Reinhold Co., New York (1975).
119. Gill, P., and W. Murray. *Math. Prog.* **14** (1978): 349.
120. Gill, P. E., W. Murray, and M. Wright. *Practical Optimization,* Academic, New York (1981).
121. Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning,* Addison-Wesley (1989).
122. Goldfarb, D. *Math. Comp.* **24** (1970): 23.
123. ——— and A. Idnani. *Math. Prog.* **27** (1983): 1.
124. ——— and M. J. Todd. "Linear Programming," Chapter II in *Optimization* (G. L. Nemhauser, A. H. G. Rinnoy Kan, and M. J. Todd, eds.), North Holland, Amsterdam (1989).
125. Gottlieb, D., and S. A. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications,* SIAM, Philadelphia (1977).
126. Gradshteyn, I. S., and I. M. Ryzhik. *Tables of Integrals, Series, and Products,* Academic, New York (1980).
127. Greenberg, M. M. *Foundations of Applied Mathematics,* Prentice Hall, Englewood Cliffs, NJ (1978).
128. Groetsch, C. W. *Generalized Inverses of Linear Operators,* Marcel Dekker, New York (1977).
129. ———. *Elements of Applicable Functional Analysis,* Marcel Dekker, New York (1980).
130. Gunzburger, M. D. *Finite Element Methods for Viscous Incompressible Flows,* Academic, New York (1989).
131. Gustafson, R. D., and P. D. Frisk. *Plane Trigonometry,* Brooks/Cole, Monterey, CA (1982).
132. Haberman, R. *Mathematical Models,* Prentice Hall, Englewood Cliffs, NJ (1977).
133. Hageman, L. A., and D. M. Young. *Applied Iterative Methods,* Academic, New York (1981).
134. Hamburg, M. *Statistical Analysis for Decision Making,* 2d ed., Harcourt, New York (1977).
135. Hamming, R. W. *Numerical Methods for Scientists and Engineers,* 2d ed., McGraw-Hill, New York (1973).
136. Han, S-P. *J. Opt. Theo. Applics.* **22** (1977): 297.
137. Hanna, R. *Fourier Series and Integrals of Boundary Value Problems,* Wiley, New York (1982).
138. Hanna, O. T., and O. C. Sandall. *Computational Methods in Chemical Engineering,* Prentice Hall, Upper Saddle River, NJ (1994).
139. Hardy, G. H., J. E. Littlewood, and G. Polya. *Inequalities,* 2d ed., Cambridge University Press, Cambridge (1952).
140. Haykin, S. *Neural Networks: A Comprehensive Foundation,* Macmillan, New York (1994).
141. Henrici, P. *Applied and Computational Complex Analysis,* Wiley, New York (1974).
142. Hestenes, M. R. *Conjugate Gradient Methods in Optimization,* Springer-Verlag (1980).
143. Hildebrand, F. B. *Introduction to Numerical Analysis,* 2d ed., McGraw-Hill, New York (1974).
144. ———. *Advanced Calculus for Applications,* 2d ed., Prentice Hall, Englewood Cliffs, NJ (1976).
145. Hill, J. M. *Differential Equations and Group Methods for Scientists and Engineers,* CRC Press (1992).
146. Hille, E. *Ordinary Differential Equations in the Complex Domain,* Wiley (1976).
147. Hille, E. *Methods in Classical and Functional Analysis,* Addison-Wesley, Reading, MA (1972).
148. Hindmarsh, A. C. *ACM SIGNUM Newsletter* **15** (1980): 10–11.
149. Hindmarsh, A. C. "GEARB: Solution of Ordinary Differential Equations Having Banded Jacobian," UCID-30059, Rev. 1 Computer Documentation, Lawrence Livermore Laboratory, University of California (1975).
150. Hornbeck, R. W. *Numerical Methods,* Prentice Hall, Englewood Cliffs, NJ (1975).
151. Hougen, O. A., R. M. Watson, and R. A. Ragatz. Part II, "Thermodynamics," in *Chemical Process Principles,* 2d ed., Wiley, New York (1959).
152. Householder, A. S. *The Theory of Matrices in Numerical Analysis,* Dover, New York (1979).

153. ———. *Numerical Treatment of a Single Nonlinear Equation,* McGraw-Hill, New York, (1970) and Dover, New York (1980).
154. Houstis, E. N., W. F. Mitchell, and T. S. Papatheodoros. *Int. J. Num. Methods Engn.* **19** (1983): 665–704.
155. Isaacson, E., and H. B. Keller. *Analysis of Numerical Methods,* Wiley, New York (1966).
156. Jeffreys, H., and B. Jeffreys. *Methods of Mathematical Physics,* 3d ed., Cambridge University Press, London (1972).
157. Jennings, A., and J. J. McKeown. *Matrix Computations for Engineers and Scientists,* Wiley, New York (1992).
158. Johnson, R. E., and F. L. Kiokemeister. *Calculus with Analytic Geometry,* 4th ed., Allyn and Bacon, Boston (1969).
159. Joseph, D. D., M. Renardy, and J. C. Saut. *Arch. Rational Mech. Anal.* **87** (1985): 213–251.
160. Juncu, G., and R. Mihail. *Comp. Chem. Eng.* **13** (1989): 259–270.
161. Kalos, M. H., and P. A. Whitlock. *Monte Carlo Methods,* vol. I, Wiley, New York (1986).
162. Kantorovich, L. V., and G. P. Akilov. *Functional Analysis,* 2d ed., Pergamon, Oxford (1982).
163. Kaplan, W. *Advanced Calculus,* 2d ed., Addison-Wesley, Reading, MA (1973).
164. Kardestuncer, H., and D. H. Norrie (ed.). *Finite Element Handbook,* McGraw-Hill (1987).
165. Karmarker, N. *Combinatorica* **4** (1984): 373–395.
166. Keedy, M. L., and M. L. Bittinger. *Trigonometry: Triangles and Functions,* 3d ed., Addison-Wesley, New York (1983).
167. Keller, H. B. *Numerical Methods for Two-Point Boundary-Value Problems,* Blaisdell, New York (1972).
168. Kemeny, J. G., J. L. Snell, and G. L. Thompson. *Introduction to Finite Mathematics,* 3d ed., Prentice Hall, Englewood Cliffs, NJ (1975).
169. Kendall, M. G., A. Stuart, J. K. Ord, and A. O'Hogan. *Advanced Theory of Statistics,* Halsted, New York (1994).
170. Kevorkian, J., and J. D. Cole: *Perturbation Methods in Applied Mathematics,* Springer-Verlag, New York (1981).
171. Kincaid, D. R., and D. M. Young. "Survey of Iterative Methods," in *Encyclopedia of Computer Science and Technology,* Marcel Dekker, New York (1979).
172. Krantz, S. G. *Function Theory of Several Complex Variables,* 2d ed., Wadsworth and Brooks, New York (1992).
173. Kreyszig, E. *Advanced Engineering Mathematics,* 7th ed., Wiley, New York (1993).
174. ———. *Introductory Functional Analysis with Applications,* Wiley, New York (1978).
175. Krieger, J. H. *Chem. Eng. News* **73** (Mar. 27, 1995): 50–61.
176. Kubicek, M., and M. Marek. *Computational Methods in Bifurcation Theory and Dissipative Structures,* Springer-Verlag, Berlin (1983).
177. Kuhn, H. W., and A. W. Tucker. "Nonlinear Programming" in Neyman, J. (ed.), *Proc. Second Berkeley Symp. Mathematical Statistics and Probability* (1951): 402–411.
178. Kuipers, B. *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge,* MIT Press, Boston (1994).
179. Kyrala, A. *Applied Functions of a Complex Variable,* Interscience, New York (1972).
180. Lagerstrom, P. A. *Matched Asymptotic Expansions: Ideas and Techniques,* Springer-Verlag (1988).
181. Lambert, J. D. *Computational Methods in Ordinary Differential Equations,* Wiley, New York (1973).
182. Lanczos, C. *J. Math. Phys.* **17** (1938): 123–199.
183. Lanczos, C. *Applied Analysis,* Prentice Hall, Englewood Cliffs, NJ (1956).
184. Lapidus, L., and G. F. Pinder. *Numerical Solution of Partial Differential Equations in Science and Engineering,* Interscience, New York (1982).
185. Lapidus, L., and J. Seinfeld. *Numerical Solution of Ordinary Differential Equations,* Academic, New York (1971).
186. Lapin, L. L. *Statistics for Modern Business Decisions,* 2d ed., Harcourt, New York (1982).
187. Lau, H. T. *A Numerical Library in C for Scientists and Engineers,* CRC Press (1995).
188. Lawrence, J. D. *A Catalog of Special Plane Curves,* Dover, New York (1972).
189. Lawson, C. L., and R. J. Hanson. *Solving Least Squares Problems,* Prentice Hall, Englewood Cliffs, NJ (1974).
190. Lebedev, N. N. *Special Functions and Their Applications,* Dover, New York (1972).
191. Leithold, L. *College Algebra and Trigonometry,* Addison-Wesley (1989).
192. LeVeque, R. J. *Numerical Methods for Conservation Laws,* Birkhäuser, Basel (1992).
193. Levy, H. *Analytic Geometry,* Harcourt, Brace & World, New York (1969).
194. Lin, C. C., and L. A. Segel. *Mathematics Applied to Deterministic Problems in the Natural Sciences,* Macmillan, New York (1974).
195. Linz, P. *Analytical and Numerical Methods for Volterra Equations,* SIAM Publications, Philadelphia (1985).
196. Liusternik, L. A., and V. J. Sobolev. *Elements of Functional Analysis,* 3d ed., Wiley, New York (1974).
197. Luke, Y. L. *Mathematical Functions and Their Applications,* Academic, New York (1975).
198. Luyben, W. L. *Process Modeling, Simulation and Control for Chemical Engineers,* 2d ed., McGraw-Hill, New York (1990).
199. MacDonald, W. B., A. N. Hrymak, and S. Treiber. "Interior Point Algorithms for Refinery Scheduling Problems" in *Proc. 4th Annual Symp. Process Systems Engineering* (Aug. 5–9, 1991): III.13.1–16.
200. Mackerle, J., and C. A. Brebbia (eds.). *Boundary Element Reference Book,* Springer Verlag, Berlin-Heidelberg, New York and Tokyo (1988).
201. Mah, R. S. H. *Chemical Process Structures and Information Flows,* Butterworths (1990).
202. Mansfield, R. *Trigonometry with Applications,* Wadsworth, New York (1972).
203. Margenau, H., and G. M. Murphy. *The Mathematics of Physics and Chemistry,* 2d ed., Van Nostrand, Princeton, NJ (1956).
204. Martin, R. H. Jr. *Ordinary Differential Equations,* McGraw-Hill, New York (1983).
205. Mavrovouniotis, Michael L. (ed.). *Artificial Intelligence in Process Engineering,* Academic, Boston (1990).
206. McIntosh, A. *Fitting Linear Models: An Application of Conjugate Gradient Algorithms,* Springer-Verlag, New York (1982).
207. McCormick, G. P. *Nonlinear Programming: Theory, Algorithms, and Applications,* Wiley, New York (1983).
208. *McGraw-Hill Encyclopedia of Science and Technology,* McGraw-Hill, New York (1971).
209. Mei, C. C. *Mathematical Analysis in Engineering,* Cambridge (1995).
210. Mitchell, A. R., and R. Wait. *The Finite Element Method in Partial Differential Equations,* Wiley, New York (1977).
211. Mood, A. M., R. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics,* 3d ed., McGraw-Hill, New York (1974).
212. Morse, P. M., and H. Feshbach. *Methods of Theoretical Physics,* vols. I and II, McGraw-Hill, New York (1953).
213. Morton, K. W., and D. F. Mayers. *Numerical Solution of Partial Differential Equations,* Cambridge (1995).
214. Nayfeh, A. H. *Perturbation Methods,* Wiley, New York (1973).
215. ———. *Introduction to Perturbation Techniques,* Wiley, New York (1981).
216. Naylor, A. W., and G. R. Sell. *Linear Operator Theory in Engineering and Science,* Springer-Verlag, New York (1982).
217. Oberhettinger, F. *Fourier Expansions: A Collection of Formulas,* Academic, New York (1973).
218. Ogunnaike, B. A., and W. H. Ray. *Process Dynamics, Modeling, and Control,* Oxford University Press (1994).
219. Ortega, J. M. *Numerical Analysis: A Second Course,* SIAM (1990).
220. Pao, C. V. *Nonlinear Parabolic and Elliptic Equations,* Plenum (1992).
221. Peaceman, D. W. *Fundamentals of Numerical Reservoir Simulation,* Elsevier, Amsterdam (1977).
222. Pearson, Carl E. (ed.). *Handbook of Applied Mathematics,* 2d ed., Van Nostrand Reinhold Co., New York (1983).
223. Perlmutter, D. *Stability of Chemical Reactors,* Prentice Hall, Englewood Cliffs, NJ (1972).
224. Petzold, L. R. "A Description of DASSL: A Differential-Algebraic System Solver," Sandia National Laboratory Report SAND82-8637; also in Stepleman, R. S. et al., eds. *IMACS Trans. on Scientific Computing,* vol. 1, pp. 65–68.
225. Pike, R. W. *Optimization for Engineering Systems,* Van Nostrand Reinhold (1986).
226. Pontelides, C. C., D. Gritsis, K. R. Morison, and R. W. H. Sargent. *Comp. Chem. Eng.* **12** (1988): 449–454.
227. Poulain, C. A., and B. A. Finlayson. *Int. J. Num. Methods Fluids* **17** (1993): 839–859.
228. Powell, M. J. D. "A Fast Algorithm for Nonlinearly Constrained Optimization Calculations," *Lecture Notes in Mathematics* **630** (1977).
229. Powers, D. L. *Boundary Value Problems,* Academic, New York (1972).
230. Prenter, P. M. *Splines and Variational Methods,* Wiley, New York (1975).
231. Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes,* Cambridge University Press, Cambridge (1986).
232. Quantrille, T. E., and Y. A. Liu. *Artificial Intelligence in Chemical Engineering,* Academic Press, San Diego (1991).
233. Quarteroni, A., and A. Valli: *Numerical Approximation of Partial Differential Equations,* Springer-Verlag (1994).
234. Råde, L., and B. Westergren. β *Mathematics Handbook,* 2d ed., Chartwell-Bratt, Lund, Sweden (1990).
235. Rainville, E. D. *Special Functions,* Chelsea Publishing Company, New York (1972).
236. Rainville, E. D., and P. E. Bedient. *Elementary Differential Equations,* 7th ed., Macmillan, New York (1989).

237. Rall, L. B. *Computational Solution of Nonlinear Operator Equations,* Wiley, New York (1969) and Dover, New York (1981).
238. Ralston, A., and A. Rabinowitz. *A First Course in Numerical Analysis,* 2d ed., McGraw-Hill, New York (1978).
239. Ramirez, W. F. *Computational Methods for Process Simulations,* Butterworths, Boston (1989).
240. Rauch, J. *Partial Differential Equations,* Springer-Verlag (1991).
241. Reddy, J. N., and D. K. Gartling. *The Finite Element Method in Heat Transfer and Fluid Dynamics,* CRC Press (1994).
242. Reklaitis, G. V. *Introduction to Material and Energy Balances,* Wiley (1983).
243. Rekalitis, G. V., and H. D. Spriggs. *Proceedings of the First International Conference on Foundations of Computer-Aided Operations,* Elsevier Science Publishers, Inc., New York (1987).
244. Reklaitis, G. V., A. Ravindran, and K. M. Ragsdell. *Engineering Optimization Methods and Applications,* Wiley, New York (1983).
245. Rhee, H.-K., R. Aris, and N. R. Amundson. *First-Order Partial Differential Equations,* vol. I., Prentice Hall, Englewood Cliffs, NJ (1986).
246. ———. *Matrix Computations and Mathematical Software,* McGraw-Hill, New York (1981).
247. ———. *Numerical Methods, Software, and Analysis,* 2d ed., Academic, New York (1993).
248. Rich, E., and K. Kevin. *Artificial Intelligence,* 2d ed., McGraw-Hill, New York (1991).
249. Riggs, J. B. *An Introduction to Numerical Methods for Chemical Engineers,* Texas Tech Univ. Press, Lubbock, TX (1994).
250. Rippin, D. W. T., J. C. Hale, and J. F. Davis (ed.). *Proceedings of the Second International Conference on Foundations of Computer-Aided Operations,* CACHE Corporation, Austin, TX (1994).
251. Ritchmyer, R., and K. Morton. *Difference Methods for Initial-Value Problems,* 2d ed., Interscience, New York (1967).
252. Saaty, T. L., and J. Bram. *Nonlinear Mathematics,* McGraw-Hill, New York (1964) and Dover, New York (1981).
253. Schiesser, W. E. *The Numerical Method of Lines,* Academic Press (1991).
254. Schittkowski, K. *Num. Math.* **38** (1982): 83.
255. Seader, J. D. *Computer Modeling of Chemical Processes,* AIChE Monog. Ser. No. 15 (1985).
256. Seborg, D. E., T. F. Edgar, and D. A. Mellichamp. *Process Dynamics and Control,* Wiley, New York (1989).
257. Shampine, L. *Numerical Solution of Ordinary Differential Equations,* Chapman & Hall (1994).
258. Shapiro, S. C., D. Eckroth et. al (ed.). *Encyclopedia of Artificial Intelligence,* Wiley, New York (1987).
259. Shanno, D. F. *Math. Comp.* **24** (1970): 647.
260. Shenk, A. *Calculus and Analytic Geometry,* Goodyear Publishing Company, Santa Monica, CA (1977).
261. Shockley, J. E. *Calculus and Analytic Geometry,* Saunders, Philadelphia (1982).
262. Siirola, J. J., I. E. Grossmann, and G. Stephanopoulos. *Proceedings of the Second International Conference on Foundations of Computer-Aided Design,* Elsevier Science Publishers, Inc., New York (1990).
263. Simmons, G. F. *Differential Equations,* McGraw-Hill, New York (1972).
264. Simmonds, J. G. *A Brief on Tensor Analysis,* Springer-Verlag (1994).
265. Sincich, T., and Mendenhall, W. *Statistics for Engineering and the Sciences,* 4th ed., Prentice Hall, Englewood Cliffs, NJ (1995).
266. Sincovec, R. F. *Math. Comp.* **26** (1972): 893–895.
267. Smith, I. M., J. L. Siemienivich, and I. Gladwell. "A Comparison of Old and New Methods for Large Systems of Ordinary Differential Equations Arising from Parabolic Partial Differential Equations," Num. Anal. Rep. Department of Engineering, no. 13, University of Manchester, England (1975).
268. Smith, W. K. *Analytic Geometry,* Macmillan (1972).
269. Sobel, M. A., and N. Lerner. *College Algebra,* Prentice Hall, Englewood Cliffs, NJ (1983).
270. Sod, G. A. *Numerical Methods in Fluid Dynamics,* Cambridge Univ. Press (1985).
271. Sokolnikoff, I. S., and Sokolnikoff, E. S. *Higher Mathematics for Engineers and Physicists,* McGraw-Hill, New York (1941).
272. Spiegel, M. R. *Applied Differential Equations,* 3d ed., Prentice Hall, Englewood Cliffs, NJ (1981).
273. Stakgold, I. *Green's Functions and Boundary Value Problems,* Interscience, New York (1979).
274. Stein, S. K. *Calculus and Analytic Geometry,* 3d ed., McGraw-Hill, New York (1982).
275. Stephanopoulos, G., and J. F. Davis (eds.). *Artificial Intelligence in Process Engineering,* CACHE Monograph Series, CACHE, Austin (1990–1992).
276. Stephanopoulos, G., and H. Chonghun. "Intelligent Systems in Process Engineering: A Review," *Proceedings of PSE '94,* Korea (1994).
277. Stillwell, J. C. *Elements of Algebra,* CRC Press, New York (1994).
278. Stoer, J., and R. Bulirsch. *Introduction to Numerical Analysis,* Springer, New York (1993).
279. Strang, G. *Linear Algebra and Its Applications,* 2d ed., Academic, New York (1980).
280. Strang, G. *Introduction to Linear Algebra,* Wellesley-Cambridge, Cambridge, MA (1993).
281. ——— and G. Fix. *An Analysis of the Finite Element Method,* Prentice Hall, Englewood Cliffs, NJ (1973).
282. Swokowski, E. W. *Calculus with Analytic Geometry,* 2d ed., Prindle, Weber & Schmidt, Boston (1981).
283. Taylor, A. E., and D. C. Lay. *Introduction to Functional Analysis,* 2d ed., Wiley, New York (1980).
284. Umeda, T., and A. Ichikawa. *I&EC Proc. Design Develop.* **10** (1971): 229.
285. Vasantharajan, S., and L. T. Biegler. *Computers and Chemical Engineering* **12** (1988): 1087.
286. Vemuri, V., and W. Karplus. *Digital Computer Treatment of Partial Differential Equations,* Prentice Hall, Englewood Cliffs, NJ (1981).
287. Vichnevetsky, R. *Computer Methods for Partial Differential Equations,* vols. 1 and 2, Prentice Hall, Englewood Cliffs, NJ (1981, 1982).
288. Villadsen, J. V., and M. L. Michelsen. *Solution of Differential Equation Models by Polynomial Approximation,* Prentice Hall, Englewood Cliffs, NJ (1978).
289. Villadsen, J., and W. E. Stewart. *Chem. Eng. Sci.* **22** (1967): 1483–1501.
290. Walas, S. M. *Modeling with Differential Equations in Chemical Engineering,* Butterworth-Heinemann, Stoneham, MA (1991).
291. Weisberg, S. *Applied Linear Regression,* 2d ed., Wiley, New York (1985).
292. Weld, D. S., and J. de Kleer (ed.). *Readings in Qualitative Reasoning About Physical Systems,* Morgan Kaufman, San Mateo, CA (1990).
293. Westerberg, A. W., H. P. Hutchison, R. L. Motard, and P. Winter. *Process Flowsheeting,* Cambridge University Press, London (1979).
294. Westerberg, A. W., and H. H. Chien (ed.). *Proceedings of the Second International Conference on Foundations of Computer-Aided Design,* CACHE Corporation, Austin, TX (1984).
295. Westerberg, A. W. "Optimization" in A. K. Sunol, D. W. T. Rippin, G. V. Reklaitis, O. Hortacsu (eds.), *Batch Processing Systems Engineering: Current Status and Future Directions,* vol. 143, NATO ASI Series F, Springer, Berlin (1995).
296. Whipkey, K. L., and M. N. Whipkey. *The Power of Calculus,* 3d ed., Wiley, New York (1979).
297. Wilkinson, J. H. *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford (1988).
298. Williams, G. *Computational Linear Algebra with Models,* 2d ed., Allyn and Bacon, Boston (1981).
299. Wolfram, S. *Mathematica,* Addison-Wesley, New York (1988).
300. Wouk, A. *A Course of Applied Functional Analysis,* Interscience, New York (1979).
301. Wylie, C. R. *Advanced Engineering Mathematics,* 5th ed., McGraw-Hill, New York (1982).
302. Young, D. M. *Iterative Solution for Large Linear Systems,* Academic, New York (1971).
303. Zienkiewicz, O. C., and R. L. Taylor. *The Finite Element Method,* McGraw-Hill, London (1989).
304. ——— and K. Morgan. *Finite Elements and Approximations,* Wiley, New York (1983).

References for General and Specific Topics

Advanced engineering mathematics:
Upper undergraduate level, 19, 80, 127, 144, 156, 173, 194, 203, 209, 301. Graduate level, 79, 127, 212, 273. Mathematical tables, mathematical dictionaries, and handbooks of mathematical functions and formulas, 1, 28, 48, 57, 118, 126, 188, 208, 217, 222, 234. Mathematical modeling of physical phenomena, 17, 19, 31, 44, 98, 132, 194, 273. Mathematical theory of reaction, diffusion, and transport processes, 10, 16, 19, 88, 223. Mathematical methods in chemical engineering, 13, 15, 61, 85, 104, 106, 138, 239, 249, 288. Inequalities, 28, 126, 139, 290.

Vector and tensor analysis, 18, 163, 173, 264.

Special functions in physics and engineering, 190, 197, 235.

Green's functions and applications, 75, 127, 273.

Perturbation and asymptotic methods in applied mathematics, 170, 215, 216.

Approximation theory and interpolation, 86, 87.

Functional analysis; linear operators, 129, 147, 162, 174, 196, 216, 226, 283, 300.

Generalized inverses and least-squares problems, 33, 128, 189.

MATHEMATICS

GENERAL

The basic problems of the sciences and engineering fall broadly into three categories:

1. *Steady state problems.* In such problems the configuration of the system is to be determined. This solution does not change with time but continues indefinitely in the same pattern, hence the name "steady state." Typical chemical engineering examples include steady temperature distributions in heat conduction, equilibrium in chemical reactions, and steady diffusion problems.
2. *Eigenvalue problems.* These are extensions of equilibrium problems in which critical values of certain parameters are to be determined in addition to the corresponding steady-state configurations. The determination of eigenvalues may also arise in propagation problems. Typical chemical engineering problems include those in heat transfer and resonance in which certain boundary conditions are prescribed.
3. *Propagation problems.* These problems are concerned with predicting the subsequent behavior of a system from a knowledge of the initial state. For this reason they are often called the transient (time-varying) or unsteady-state phenomena. Chemical engineering examples include the transient state of chemical reactions (kinetics), the propagation of pressure waves in a fluid, transient behavior of an adsorption column, and the rate of approach to equilibrium of a packed distillation column.

The mathematical treatment of engineering problems involves four basic steps:

1. *Formulation.* The expression of the problem in mathematical language. That translation is based on the appropriate physical laws governing the process.
2. *Solution.* Appropriate mathematical operations are accomplished so that logical deductions may be drawn from the mathematical model.
3. *Interpretation.* Development of relations between the mathematical results and their meaning in the physical world.
4. *Refinement.* The recycling of the procedure to obtain better predictions as indicated by experimental checks.

Steps 1 and 2 are of primary interest here. The actual details are left to the various subsections, and only general approaches will be discussed.

The formulation step may result in algebraic equations, difference equations, differential equations, integral equations, or combinations of these. In any event these mathematical models usually arise from statements of physical laws such as the laws of mass and energy conservation in the form.

Input of conserved quantity − output of conserved quantity + conserved quantity produced = accumulation of conserved quantity

Rate of input of conserved quantity − rate of output of conserved quantity + rate of conserved quantity produced = rate of accumulation of conserved quantity

These statements may be abbreviated by the statement

Input − output + production = accumulation

When the basic physical laws are expressed in this form, the formulation is greatly facilitated. These expressions are quite often given the names, "material balance," "energy balance," and so forth. To be a little more specific, one could write the law of conservation of energy in the steady state as

Rate of energy in − rate of energy out + rate of energy produced = 0

Many general laws of the physical universe are expressible by differential equations. Specific phenomena are then singled out from the infinity of solutions of these equations by assigning the individual initial or boundary conditions which characterize the given problem. In mathematical language one such problem, the equilibrium problem, is called a boundary-value problem (Fig. 3-1). Schematically, the problem is characterized by a differential equation plus an open region in which the equation holds and, on the boundaries of the region, by certain conditions (boundary conditions) that are dictated by the physical problem. The solution of the equation must satisfy the differential equation inside the region and the prescribed conditions on the boundary.

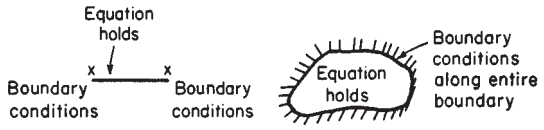

FIG. 3-1 Boundary conditions.

In mathematical language, the propagation problem is known as an initial-value problem (Fig. 3-2). Schematically, the problem is characterized by a differential equation plus an open region in which the equation holds. The solution of the differential equation must satisfy the initial conditions plus any "side" boundary conditions.

The description of phenomena in a "continuous" medium such as a gas or a fluid often leads to partial differential equations. In particular, phenomena of "wave" propagation are described by a class of partial differential equations called "hyperbolic," and these are essentially different in their properties from other classes such as those that describe equilibrium ("elliptic") or diffusion and heat transfer ("parabolic"). Prototypes are:

1. *Elliptic.* Laplace's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

Poisson's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = g(x,y)$$

These do not contain the variable t (time) explicitly; accordingly, their solutions represent equilibrium configurations. Laplace's equation corresponds to a "natural" equilibrium, while Poisson's equation corresponds to an equilibrium under the influence of an external force of density proportional to $g(x, y)$.

2. *Parabolic.* The heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

describes nonequilibrium or propagation states of diffusion as well as heat transfer.

3. *Hyperbolic.* The wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

describes wave propagation of all types when the assumption is made that the wave amplitude is small and that interactions are linear.

The solution phase has been characterized in the past by a concentration on methods to obtain analytic solutions to the mathematical

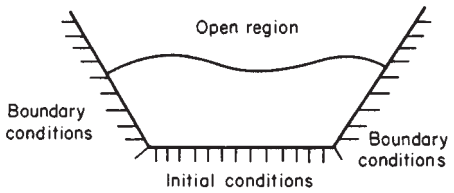

FIG. 3-2 Propagation problem.

equations. These efforts have been most fruitful in the area of the linear equations such as those just given. However, many natural phenomena are nonlinear. While there are a few nonlinear problems that can be solved analytically, most cannot. In those cases, numerical methods are used. Due to the widespread availability of software for computers, the engineer has quite good tools available.

Numerical methods almost never fail to provide an answer to any particular situation, but they can never furnish a general solution of any problem.

The mathematical details outlined here include both analytic and numerical techniques useful in obtaining solutions to problems.

Our discussion to this point has been confined to those areas in which the governing laws are well known. However, in many areas, information on the governing laws is lacking. Interest in the application of statistical methods to all types of problems has grown rapidly since World War II. Broadly speaking, statistical methods may be of use whenever conclusions are to be drawn or decisions made on the basis of experimental evidence. Since statistics could be defined as the technology of the scientific method, it is primarily concerned with the first two aspects of the method, namely, the performance of experiments and the drawing of conclusions from experiments. Traditionally the field is divided into two areas:

1. *Design of experiments.* When conclusions are to be drawn or decisions made on the basis of experimental evidence, statistical techniques are most useful when experimental data are subject to errors. The design of experiments may then often be carried out in such a fashion as to avoid some of the sources of experimental error and make the necessary allowances for that portion which is unavoidable. Second, the results can be presented in terms of probability statements which express the reliability of the results. Third, a statistical approach frequently forces a more thorough evaluation of the experimental aims and leads to a more definitive experiment than would otherwise have been performed.

2. *Statistical inference.* The broad problem of statistical inference is to provide measures of the uncertainty of conclusions drawn from experimental data. This area uses the theory of probability, enabling scientists to assess the reliability of their conclusions in terms of probability statements.

Both of these areas, the mathematical and the statistical, are intimately intertwined when applied to any given situation. The methods of one are often combined with the other. And both in order to be successfully used must result in the numerical answer to a problem—that is, they constitute the means to an end. Increasingly the numerical answer is being obtained from the mathematics with the aid of computers.

MISCELLANEOUS MATHEMATICAL CONSTANTS

Numerical values of the constants that follow are approximate to the number of significant digits given.

$\pi = 3.1415926536$	Pi
$e = 2.7182818285$	Napierian (natural) logarithm base
$\gamma = 0.5772156649$	Euler's constant
$\ln \pi = 1.1447298858$	Napierian (natural) logarithm of pi, base e
$\log \pi = 0.4971498727$	Briggsian (common logarithm of pi, base 10
Radian = 57.2957795131°	
Degree = 0.0174532925 rad	
Minute = 0.0002908882 rad	
Second = 0.0000048481 rad	

THE REAL-NUMBER SYSTEM

The natural numbers, or counting numbers, are the positive integers: 1, 2, 3, 4, 5, ... The negative integers are –1, –2, –3, ...

A number in the form a/b, where a and b are integers, $b \neq 0$, is a rational number. A real number that cannot be written as the quotient of two integers is called an irrational number, e.g., $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, π, e, $\sqrt[3]{2}$.

There is a one-to-one correspondence between the set of real numbers and the set of points on an infinite line (coordinate line).

Order among Real Numbers; Inequalities

$a > b$ means that $a - b$ is a positive real number.
If $a < b$ and $b < c$, then $a < c$.
If $a < b$, then $a \pm c < b \pm c$ for any real number c.
If $a < b$ and $c > 0$, then $ac < bc$.
If $a < b$ and $c < 0$, then $ac > bc$.
If $a < b$ and $c < d$, then $a + c < b + d$.
If $0 < a < b$ and $0 < c < d$, then $ac < bd$.
If $a < b$ and $ab > 0$, then $1/a > 1/b$.
If $a < b$ and $ab < 0$, then $1/a < 1/b$.

Absolute Value For any real number x, $|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$

Properties

If $|x| = a$, where $a > 0$, then $x = a$ or $x = -a$.
$|x| = |-x|$; $-|x| \leq x \leq |x|$; $|xy| = |x|\,|y|$.
If $|x| < c$, then $-c < x < c$, where $c > 0$.
$||x| - |y|| \leq |x + y| \leq |x| + |y|$.
$\sqrt{x^2} = |x|$.

Proportions If $\frac{a}{b} = \frac{c}{d}$, then $\frac{a+b}{b} = \frac{c+d}{d}$, $\frac{a-b}{b} = \frac{c-d}{d}$, $\frac{a-b}{a+b} = \frac{c-d}{c+d}$.

Indeterminants

Form	Example	
$(\infty)(0)$	xe^{-x}	$x \to \infty$
0^0	x^x	$x \to 0^+$
∞^0	$(\tan x)^{\cos x}$	$x \to \frac{1}{2}\pi^-$
1^∞	$(1+x)^{1/x}$	$x \to 0^+$
$\infty - \infty$	$\sqrt{x+1} - \sqrt{x-1}$	$x \to \infty$
$\frac{0}{0}$	$\frac{\sin x}{x}$	$x \to 0$
$\frac{\infty}{\infty}$	$\frac{e^x}{x}$	$x \to \infty$

Limits of the type $0/\infty$, $\infty/0$, 0^∞, $\infty \cdot \infty$, $(+\infty) + (+\infty)$, and $(-\infty) + (-\infty)$ are *not* indeterminate forms.

Integral Exponents (Powers and Roots) If m and n are positive integers and a, b are numbers or functions, then the following properties hold:

$$a^{-n} = 1/a^n \qquad a \neq 0$$
$$(ab)^n = a^n b^n$$
$$(a^n)^m = a^{nm}, \qquad a^n a^m = a^{n+m}$$
$$\sqrt[n]{a} = a^{1/n} \quad \text{if} \quad a > 0$$
$$\sqrt[m]{\sqrt[n]{a}} = \sqrt[mn]{a},\ a > 0$$
$$a^{m/n} = (a^m)^{1/n} = \sqrt[n]{a^m},\ a > 0$$
$$a^0 = 1\ (a \neq 0)$$
$$0^a = 0\ (a \neq 0)$$

Infinity (∞) is not a real number. It is possible to extend the real-number system by adjointing to it "∞" and "$-\infty$," and within the extended system, certain operations involving $+\infty$ or $-\infty$ are possible. For example, if $0 < a < 1$, then $a^\infty = \lim_{x \to \infty} a^x = 0$, whereas if $a > 1$, then $a^\infty = \infty$, $\infty^a = \infty$ $(a > 0)$, $\infty^a = 0$ $(a < 0)$.

Care should be taken in the case of roots and fractional powers of a product; e.g., $\sqrt{xy} \neq \sqrt{x}\sqrt{y}$ if x and y are negative. This rule applies if one is careful about the domain of the functions involved; so $\sqrt{xy} = \sqrt{x}\sqrt{y}$ if $x > 0$ and $y > 0$.

Given any number $b > 0$, there is a unique function $f(x)$ defined for all real numbers x such that (1) $f(x) = b^x$ for all rational x; (2) f is increasing if $b > 1$, constant if $b = 1$, and decreasing if $0 < b < 1$. This function is called the exponential function b^x. For any $b > 0$, $f(x) = b^x$ is

a continuous function. Also with $a,b > 0$ and x,y any real numbers, we have

$$(ab)^x = a^x b^x$$
$$b^x b^y = b^{x+y}$$
$$(b^x)^y = b^{xy}$$

The exponential function with base b can also be defined as the inverse of the logarithmic function. The most common exponential function in applications corresponds to choosing b the transcendental number e.

Logarithms
$$\log ab = \log a + \log b,\ a > 0, b > 0$$
$$\log a^n = n \log a$$
$$\log (a/b) = \log a - \log b$$
$$\log \sqrt[n]{a} = (1/n) \log a$$

The common logarithm (base 10) is denoted $\log a$ or $\log_{10} a$. The natural logarithm (base e) is denoted $\ln a$ (or in some texts $\log_e a$).

Roots If a is a real number, n is a positive integer, then x is called the nth root of a if $x^n = a$. The number of nth roots is n, but not all of them are necessarily real. The principal nth root means the following: (1) if $a > 0$ the principal nth root is the unique positive root, (2) if $a < 0$, and n odd, it is the unique negative root, and (3) if $a < 0$ and n even, it is any of the complex roots. In cases (1) and (2), the root can be found on a calculator by taking $y = \ln a/n$ and then $x = e^y$. In case (3), see the section on complex variables.

PROGRESSIONS

Arithmetic Progression

$$\sum_{k=0}^{n-1} (a + kd) = na + \frac{1}{2} n(n-1)d = \frac{n}{2}(a + \ell)$$

where ℓ is the last term, $\ell = a + (n-1)d$.

Geometric Progression

$$\sum_{k=1}^{n} ar^{k-1} = \frac{a(r^n - 1)}{r - 1} \qquad (r \neq 1)$$

Arithmetic-Geometric Progression

$$\sum_{k=0}^{n-1} (a + kd)r^k = \frac{a - [a + (n-1)d]r^n}{1 - r} + \frac{dr(1 - r^{n-1})}{(1 - r)^2} \qquad (r \neq 1)$$

$$\sum_{k=1}^{n} k^5 = \frac{1}{12} n^2(n+1)^2(2n^2 + 2n - 1)$$

$$\sum_{k=1}^{n} (2k - 1) = n^2$$

$$\sum_{k=1}^{n} (2k-1)^2 = \frac{1}{3} n(4n^2 - 1)$$

$$\sum_{k=1}^{n} (2k-1)^3 = n^2(2n^2 - 1)$$

$$\gamma = \lim_{n \to \infty} \left(\sum_{m=1}^{n} \frac{1}{m} - \ln n \right) = 0.577215$$

ALGEBRAIC INEQUALITIES

Arithmetic-Geometric Inequality Let A_n and G_n denote respectively the arithmetic and the geometric means of a set of positive numbers $a_1, a_2, \ldots, a_n$. The $A_n \geq G_n$, i.e.,

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq (a_1 a_2 \cdots a_n)^{1/n}$$

The equality holds only if all of the numbers a_i are equal.

Carleman's Inequality The arithmetic and geometric means just defined satisfy the inequality

$$\sum_{r=1}^{n} G_r \leq neA_n$$

or, equivalently,

$$\sum_{r=1}^{n} (a_1 a_2 \cdots a_r)^{1/r} \leq neA_n$$

where e is the best possible constant in this inequality.

Cauchy-Schwarz Inequality Let $a = (a_1, a_2, \ldots, a_n)$, $b = (b_1, b_2, \ldots, b_n)$, where the a_i's and b_i's are real or complex numbers. Then

$$\left| \sum_{k=1}^{n} a_k \bar{b}_k \right|^2 \leq \left(\sum_{k=1}^{n} |a_k|^2 \right) \left(\sum_{k=1}^{n} |b_k|^2 \right)$$

The equality holds if, and only if, the vectors a, b are linearly dependent (i.e., one vector is scalar times the other vector).

Minkowski's Inequality Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be any two sets of complex numbers. Then for any real number $p > 1$,

$$\left(\sum_{k=1}^{n} |a_k + b_k|^p \right)^{1/p} \leq \left(\sum_{k=1}^{n} |a_k|^p \right)^{1/p} + \left(\sum_{k=1}^{n} |b_k|^p \right)^{1/p}$$

Hölder's Inequality Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be any two sets of complex numbers, and let p and q be positive numbers with $1/p + 1/q = 1$. Then

$$\left| \sum_{k=1}^{n} a_k \bar{b}_k \right| \leq \left(\sum_{k=1}^{n} |a_k|^p \right)^{1/p} \left(\sum_{k=1}^{n} |b_k|^q \right)^{1/q}$$

The equality holds if, and only if, the sequences $|a_1|^p, |a_2|^p, \ldots, |a_n|^p$ and $|b_1|^q, |b_2|^q, \ldots, |b_n|^q$ are proportional and the argument (angle) of the complex numbers $a_k \bar{b}_k$ is independent of k. This last condition is of course automatically satisfied if $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are positive numbers.

Lagrange's Inequality Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be real numbers. Then

$$\left(\sum_{k=1}^{n} a_k b_k \right)^2 = \left(\sum_{k=1}^{n} a_k^2 \right) \left(\sum_{k=1}^{n} b_k^2 \right) - \sum_{1 \leq k \leq j \leq n} (a_k b_j - a_j b_k)^2$$

Example Two chemical engineers, John and Mary, purchase stock in the same company at times $t_1, t_2, \ldots, t_n$, when the price per share is respectively $p_1, p_2, \ldots, p_n$. Their methods of investment are different, however: John purchases x shares each time, whereas Mary invests P dollars each time (fractional shares can be purchased). Who is doing better?

While one can argue intuitively that the average cost per share for Mary does not exceed that for John, we illustrate a mathematical proof using inequalities. The average cost per share for John is equal to

$$\frac{\text{Total money invested}}{\text{Number of shares purchased}} = \frac{x \sum_{i=1}^{n} p_i}{nx} = \frac{1}{n} \sum_{i=1}^{n} p_i$$

The average cost per share for Mary is

$$\frac{nP}{\sum_{i=1}^{n} \frac{P}{p_i}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{p_i}}$$

Thus the average cost per share for John is the arithmetic mean of $p_1, p_2, \ldots, p_n$, whereas that for Mary is the harmonic mean of these n numbers. Since the harmonic mean is less than or equal to the arithmetic mean for any set of positive numbers and the two means are equal only if $p_1 = p_2 = \cdots = p_n$, we conclude that the average cost per share for Mary is less than that for John if two of the prices p_i are distinct. One can also give a proof based on the Cauchy-Schwarz inequality. To this end, define the vectors

$$a = (p_1^{-1/2}, p_2^{-1/2}, \ldots, p_n^{-1/2}) \qquad b = (p_1^{1/2}, p_2^{1/2}, \ldots, p_n^{1/2})$$

Then $a \cdot b = 1 + \cdots + 1 = n$, and so by the Cauchy-Schwarz inequality

$$(a \cdot b)^2 = n^2 \leq \sum_{i=1}^{n} \frac{1}{p_i} \sum_{i=1}^{n} p_i$$

with the equality holding only if $p_1 = p_2 = \cdots = p_n$. Therefore

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{p_i}} \leq \frac{\sum_{i=1}^{n} p_i}{n}$$

MENSURATION FORMULAS

Let A denote areas and V, volumes, in the following.

PLANE GEOMETRIC FIGURES WITH STRAIGHT BOUNDARIES

Triangles (see also "Plane Trigonometry") $A = \frac{1}{2}bh$ where b = base, h = altitude.

Rectangle $A = ab$ where a and b are the lengths of the sides.

Parallelogram (opposite sides parallel) $A = ah = ab \sin \alpha$ where a, b are the lengths of the sides, h the height, and α the angle between the sides. See Fig. 3-3.

Rhombus (equilateral parallelogram) $A = \frac{1}{2}ab$ where a, b are the lengths of the diagonals.

Trapezoid (four sides, two parallel) $A = \frac{1}{2}(a + b)h$ where the lengths of the parallel sides are a and b, and h = height.

Quadrilateral (four-sided) $A = \frac{1}{2}ab \sin \theta$ where a, b are the lengths of the diagonals and the acute angle between them is θ.

Regular Polygon of *n* Sides See Fig. 3-4.

$$A = \frac{1}{4} nl^2 \cot \frac{180°}{n}$$ where l = length of each side

$$R = \frac{l}{2} \csc \frac{180°}{n}$$ where R is the radius of the circumscribed circle

$$r = \frac{l}{2} \cot \frac{180°}{n}$$ where r is the radius of the inscribed circle

$$\beta = \frac{360°}{n}$$

$$\theta = \frac{(n-2)180°}{n}$$

$$l = 2r \tan \frac{\beta}{2} = 2R \sin \frac{\beta}{2}$$

Inscribed and Circumscribed Circles with Regular Polygon of* n *Sides Let l = length of one side.

Figure	n	Area	Radius of circumscribed circle	Radius of inscribed circle
Equilateral triangle	3	$0.4330\, l^2$	$0.5774\, l$	$0.2887\, l$
Square	4	$1.0000\, l^2$	$0.7071\, l$	$0.5000\, l$
Pentagon	5	$1.7205\, l^2$	$0.8507\, l$	$0.6882\, l$
Hexagon	6	$2.5981\, l^2$	$1.0000\, l$	$0.8660\, l$
Octagon	8	$4.8284\, l^2$	$1.3065\, l$	$1.2071\, l$
Decagon	10	$7.6942\, l^2$	$1.6180\, l$	$1.5388\, l$
Dodecagon	12	$11.1962\, l^2$	$1.8660\, l$	$1.9318\, l$

***Radius* r *of Circle Inscribed in Triangle with Sides* a, b, c**

$$r = \sqrt{\frac{(s-a)(s-b)(s-c)}{s}}$$ where $s = \frac{1}{2}(a + b + c)$

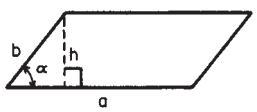

FIG. 3-3 Parallelogram.

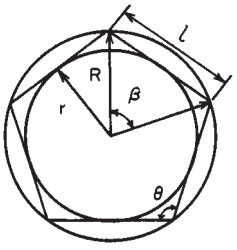

FIG. 3-4 Regular polygon.

Radius* R *of Circumscribed Circle

$$R = \frac{abc}{4\sqrt{s(s-a)(s-b)(s-c)}}$$

***Area of Regular Polygon of* n *Sides Inscribed in a Circle of Radius* r**

$$A = (nr^2/2) \sin (360°/n)$$

Perimeter of Inscribed Regular Polygon

$$P = 2nr \sin (180°/n)$$

***Area of Regular Polygon Circumscribed about a Circle of Radius* r**

$$A = nr^2 \tan (180°/n)$$

Perimeter of Circumscribed Regular Polygon

$$P = 2nr \tan \frac{180°}{n}$$

PLANE GEOMETRIC FIGURES WITH CURVED BOUNDARIES

Circle (Fig. 3-5) Let
C = circumference
r = radius
D = diameter
A = area
S = arc length subtended by θ
l = chord length subtended by θ
H = maximum rise of arc above chord, $r - H = d$
θ = central angle (rad) subtended by arc S

$C = 2\pi r = \pi D$ ($\pi = 3.14159\ldots$)

$S = r\theta = \frac{1}{2}D\theta$

$l = 2\sqrt{r^2 - d^2} = 2r \sin (\theta/2) = 2d \tan (\theta/2)$

$$d = \frac{1}{2}\sqrt{4r^2 - l^2} = \frac{1}{2} l \cot \frac{\theta}{2}$$

$$\theta = \frac{S}{r} = 2 \cos^{-1} \frac{d}{r} = 2 \sin^{-1} \frac{l}{D}$$

A (circle) $= \pi r^2 = \frac{1}{4}\pi D^2$

A (sector) $= \frac{1}{2}rS = \frac{1}{2}r^2\theta$

A (segment) $= A$ (sector) $- A$ (triangle) $= \frac{1}{2}r^2(\theta - \sin \theta)$

$$= r^2 \cos^{-1} \frac{r-H}{r} - (r-H)\sqrt{2rH - H^2}$$

Ring (area between two circles of radii r_1 and r_2) The circles need not be concentric, but one of the circles must enclose the other.

$$A = \pi(r_1 + r_2)(r_1 - r_2) \qquad r_1 > r_2$$

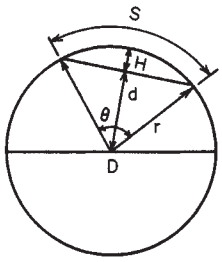

FIG. 3-5 Circle.

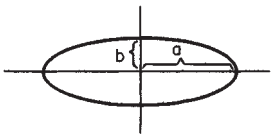

FIG. 3-6 Ellipse.

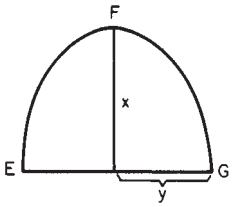

FIG. 3-7 Parabola.

Ellipse (Fig. 3-6) Let the semiaxes of the ellipse be a and b

$$A = \pi ab$$

$$C = 4aE(k)$$

where $e^2 = 1 - b^2/a^2$ and $E(e)$ is the complete elliptic integral of the second kind,

$$E(e) = \frac{\pi}{2}\left[1 - \left(\frac{1}{2}\right)^2 e^2 + \cdots\right]$$

[an approximation for the circumference $C = 2\pi\sqrt{(a^2 + b^2)/2}$].

Parabola (Fig. 3-7)

$$\text{Length of arc } EFG = \sqrt{4x^2 + y^2} + \frac{y^2}{2x} \ln \frac{2x + \sqrt{4x^2 + y^2}}{y}$$

$$\text{Area of section } EFG = \frac{4}{3}xy$$

Catenary (the curve formed by a cord of uniform weight suspended freely between two points A, B; Fig. 3-8)

$$y = a \cosh (x/a)$$

Length of arc between points A and B is equal to $2a \sinh (L/a)$. Sag of the cord is $D = a \cosh (L/a) - 1$.

SOLID GEOMETRIC FIGURES WITH PLANE BOUNDARIES

Cube Volume = a^3; total surface area = $6a^2$; diagonal = $a\sqrt{3}$, where a = length of one side of the cube.

Rectangular Parallelepiped Volume = abc; surface area = $2(ab + ac + bc)$; diagonal = $\sqrt{a^2 + b^2 + c^2}$, where a, b, c are the lengths of the sides.

Prism Volume = (area of base) × (altitude); lateral surface area = (perimeter of right section) × (lateral edge).

Pyramid Volume = ⅓ (area of base) × (altitude); lateral area of regular pyramid = ½ (perimeter of base) × (slant height) = ½ (number of sides) (length of one side) (slant height).

Frustum of Pyramid (formed from the pyramid by cutting off the top with a plane

$$V = \tfrac{1}{3}(A_1 + A_2 + \sqrt{A_1 \cdot A_2})h$$

where h = altitude and A_1, A_2 are the areas of the base; lateral area of a regular figure = ½ (sum of the perimeters of base) × (slant height).

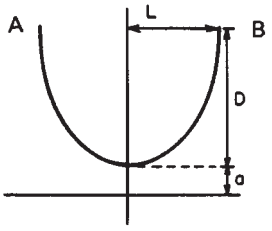

FIG. 3-8 Catenary.

Volume and Surface Area of Regular Polyhedra with Edge l

Type of surface	Name	Volume	Surface area
4 equilateral triangles	Tetrahedron	$0.1179\,l^3$	$1.7321\,l^2$
6 squares	Hexahedron (cube)	$1.0000\,l^3$	$6.0000\,l^2$
8 equilateral triangles	Octahedron	$0.4714\,l^3$	$3.4641\,l^2$
12 pentagons	Dodecahedron	$7.6631\,l^3$	$20.6458\,l^2$
20 equilateral triangles	Icosahedron	$2.1817\,l^3$	$8.6603\,l^2$

SOLIDS BOUNDED BY CURVED SURFACES

Cylinders (Fig. 3-9) V = (area of base) × (altitude); lateral surface area = (perimeter of right section) × (lateral edge).

Right Circular Cylinder $V = \pi$ (radius)^2 × (altitude); lateral surface area = 2π (radius) × (altitude).

Truncated Right Circular Cylinder

$$V = \pi r^2 h; \text{ lateral area} = 2\pi rh$$

$$h = \tfrac{1}{2}(h_1 + h_2)$$

Hollow Cylinders Volume = $\pi h(R^2 - r^2)$, where r and R are the internal and external radii and h is the height of the cylinder.

Sphere (Fig. 3-10)

$$V \text{ (sphere)} = \tfrac{4}{3}\pi R^3, \tfrac{1}{6}\pi D^3$$

$$V \text{ (spherical sector)} = \tfrac{2}{3}\pi R^2 h = \tfrac{1}{6}\pi h_1(3r_2^2 + h_1^2)$$

$$V \text{ (spherical segment of one base)} = \tfrac{1}{6}\pi h_1(3r_2^2 + h_1^2)$$

$$V \text{ (spherical segment of two bases)} = \tfrac{1}{6}\pi h_2(3r_1^2 + 3r_2^2 + h_2^2)$$

$$A \text{ (sphere)} = 4\pi R^2 = \pi D^2$$

$$A \text{ (zone)} = 2\pi Rh = \pi Dh$$

A (lune on the surface included between two great circles, the inclination of which is θ radians) = $2R^2\theta$.

Cone $V = \frac{1}{3}$ (area of base) × (altitude).

Right Circular Cone $V = (\pi/3)\,r^2 h$, where h is the altitude and r is the radius of the base; curved surface area = $\pi r\sqrt{r^2 + h^2}$, curved surface of the frustum of a right cone = $\pi(r_1 + r_2)\sqrt{h^2 + (r_1 - r_2)^2}$, where r_1, r_2 are the radii of the base and top, respectively, and h is the altitude; volume of the frustum of a right cone = $\pi(h/3)(r_1^2 + r_1 r_2 + r_2^2) = h/3(A_1 + A_2 + \sqrt{A_1 A_2})$, where A_1 = area of base and A_2 = area of top.

Ellipsoid $V = (\frac{4}{3})\,\pi abc$, where a, b, c are the lengths of the semiaxes.

Torus (obtained by rotating a circle of radius r about a line whose distance is $R > r$ from the center of the circle)

$$V = 2\pi^2 Rr^2 \qquad \text{Surface area} = 4\pi^2 Rr$$

Prolate Spheroid (formed by rotating an ellipse about its major axis [$2a$])

$$\text{Surface area} = 2\pi b^2 + 2\pi(ab/e)\sin^{-1} e \qquad V = \tfrac{4}{3}\pi ab^2$$

where a, b are the major and minor axes and e = eccentricity ($e < 1$).

Oblate Spheroid (formed by the rotation of an ellipse about its minor axis [$2b$]) Data as given previously.

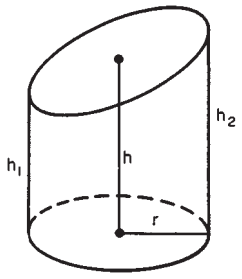

FIG. 3-9 Cylinder.

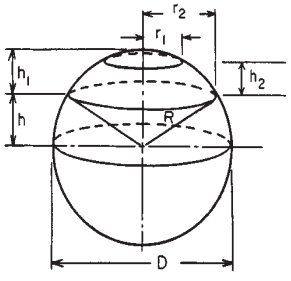

FIG. 3-10 Sphere.

$$\text{Surface area} = 2\pi a^2 + \pi \frac{b^2}{e} \ln \frac{1+e}{1-e} \qquad V = \tfrac{4}{3}\pi a^2 b$$

MISCELLANEOUS FORMULAS

See also "Differential and Integral Calculus."

Volume of a Solid Revolution (the solid generated by rotating a plane area about the x axis)

$$V = \pi \int_a^b [f(x)]^2\, dx$$

where $y = f(x)$ is the equation of the plane curve and $a \le x \le b$.

Area of a Surface of Revolution

$$S = 2\pi \int_a^b y\, ds$$

where $ds = \sqrt{1 + (dy/dx)^2}\, dx$ and $y = f(x)$ is the equation of the plane curve rotated about the x axis to generate the surface.

Area Bounded by $f(x)$, the x Axis, and the Lines $x = a$, $x = b$

$$A = \int_a^b f(x)\, dx \qquad [f(x) \ge 0]$$

Length of Arc of a Plane Curve

If $y = f(x)$,

$$\text{Length of arc } s = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx$$

If $x = g(y)$,

$$\text{Length of arc } s = \int_c^d \sqrt{1 + \left(\frac{dx}{dy}\right)^2}\, dy$$

If $x = f(t)$, $y = g(t)$,

$$\text{Length of arc } s = \int_{t_0}^{t_1} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt$$

In general, $(ds)^2 = (dx)^2 + (dy)^2$.

Theorems of Pappus (for volumes and areas of surfaces of revolution)

1. If a plane area is revolved about a line which lies in its plane but does not intersect the area, then the volume generated is equal to the product of the area and the distance traveled by the area's center of gravity.
2. If an arc of a plane curve is revolved about a line that lies in its plane but does not intersect the arc, then the surface area generated by the arc is equal to the product of the length of the arc and the distance traveled by its center of gravity.

These theorems are useful for determining volumes V and surface areas S of solids of revolution if the centers of gravity are known. If S and V are known, the centers of gravity may be determined.

IRREGULAR AREAS AND VOLUMES

Irregular Areas Let $y_0, y_1, \ldots, y_n$ be the lengths of a series of equally spaced parallel chords and h be their distance apart. The area of the figure is given approximately by any of the following:

$$A_T = (h/2)[(y_0 + y_n) + 2(y_1 + y_2 + \cdots + y_{n-1})] \qquad \text{(trapezoidal rule)}$$

$$A_s = (h/3)[(y_0 + y_n) + 4(y_1 + y_3 + y_5 + \cdots + y_{n-1}) + 2(y_2 + y_4 + \cdots + y_{n-2})] \quad (n \text{ even, Simpson's rule})$$

The greater the value of n, the greater the accuracy of approximation.

Irregular Volumes To find the volume, replace the y's by cross-sectional areas A_j and use the results in the preceding equations.

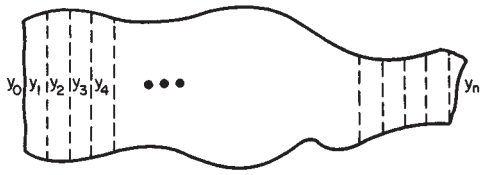

FIG. 3-11 Irregular area.

ELEMENTARY ALGEBRA

REFERENCES: 20, 102, 108, 191, 269, 277.

OPERATIONS ON ALGEBRAIC EXPRESSIONS

An algebraic expression will here be denoted as a combination of letters and numbers such as

$$3ax - 3xy + 7x^2 + 7x^{3/2} - 2.8xy$$

Addition and Subtraction Only like terms can be added or subtracted in two algebraic expressions.

Example $(3x + 4xy - x^2) + (3x^2 + 2x - 8xy) = 5x - 4xy + 2x^2$.

Example $(2^x + 3xy - 4x^{1/2}) + (3^x + 6x - 8xy) = 2^x + 3^x + 6x - 5xy - 4x^{1/2}$.

Multiplication Multiplication of algebraic expressions is term by term, and corresponding terms are combined.

Example $(2x + 3y - 2xy)(3 + 3y) = 6x + 9y + 9y^2 - 6xy^2$.

Division This operation is analogous to that in arithmetic.

Example Divide $3e^{2x} + e^x + 1$ by $e^x + 1$.

$$\begin{array}{r} \text{Dividend} \\ \text{Divisor } e^x + 1 \;\big|\; 3e^{2x} + e^x + 1 \quad 3e^x - 2 \text{ quotient} \\ \underline{3e^{2x} + 3e^x} \\ -2e^x + 1 \\ \underline{-2e^x - 2} \\ +3 \text{ (remainder)} \end{array}$$

Therefore, $3e^{2x} + e^x + 1 = (e^x + 1)(3e^x - 2) + 3$.

Operations with Zero All numerical computations (except division) can be done with zero: $a + 0 = 0 + a = a$; $a - 0 = a$; $0 - a = -a$; $(a)(0) = 0$; $a^0 = 1$ if $a \ne 0$; $0/a = 0$, $a \ne 0$. $a/0$ and $0/0$ have no meaning.

Fractional Operations

$$-\frac{x}{y} = -\left(\frac{-x}{-y}\right) = \frac{x}{-y} = \frac{-x}{y}; \quad \frac{x}{y} = \frac{-x}{-y}; \quad \frac{x}{y} = \frac{ax}{ay}, \text{ if } a \ne 0.$$

$$\frac{x}{y} \pm \frac{z}{y} = \frac{x \pm z}{y}; \quad \left(\frac{x}{y}\right)\left(\frac{z}{t}\right) = \frac{xz}{yt}; \quad \frac{x/y}{z/t} = \left(\frac{x}{y}\right)\left(\frac{t}{z}\right) = \frac{xt}{yz}$$

Factoring That process of analysis consisting of reducing a given expression into the product of two or more simpler expressions called *factors*. Some of the more common expressions are factored here:

(1) $(x^2 - y^2) = (x - y)(x + y)$

(2) $x^2 + 2xy + y^2 = (x + y)^2$

(3) $x^2 + ax + b = (x + c)(x + d)$ where $c + d = a$, $cd = b$

(4) $by^2 + cy + d = (ey + f)(gy + h)$ where $eg = b$, $fg + eh = c$, $fh = d$

(5) $x^2 + y^2 + z^2 + 2yz + 2xz + 2xy = (x + y + z)^2$

(6) $x^2 - y^2 - z^2 - 2yz = (x - y - z)(x + y + z)$

(7) $x^2 + y^2 + z^2 - 2xy - 2xz + 2yz = (x - y - z)^2$

(8) $x^3 - y^3 = (x - y)(x^2 + xy + y^2)$

(9) $(x^3 + y^3) = (x + y)(x^2 - xy + y^2)$

(10) $(x^4 - y^4) = (x - y)(x + y)(x^2 + y^2)$

(11) $x^5 + y^5 = (x + y)(x^4 - x^3y + x^2y^2 - xy^3 + y^4)$

(12) $x^n - y^n = (x - y)(x^{n-1} + x^{n-2}y + x^{n-3}y^2 + \cdots + y^{n-1})$

Laws of Exponents

$(a^n)^m = a^{nm}$; $a^{n+m} = a^n \cdot a^m$; $a^{n/m} = (a^n)^{1/m}$; $a^{n-m} = a^n/a^m$; $a^{1/m} = \sqrt[m]{a}$; $a^{1/2} = \sqrt{a}$; $\sqrt{x^2} = |x|$ (absolute value of x). For $x > 0$, $y > 0$, $\sqrt{xy} = \sqrt{x}\sqrt{y}$; for $x > 0$ $\sqrt{x^m} = x^{m/n}$; $\sqrt[n]{1/x} = 1/\sqrt{x}$

THE BINOMIAL THEOREM

If n is a positive integer,

$$(a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{3!}a^{n-3}b^3 + \cdots + b^n = \sum_{j=0}^{n}\binom{n}{j}a^{n-j}b^j$$

where $\binom{n}{j} = \frac{n!}{j!(n-j)!}$ = number of combinations of n things taken j at a time. $n! = 1 \cdot 2 \cdot 3 \cdot 4 \cdots n$, $0! = 1$.

Example Find the sixth term of $(x + 2y)^{12}$. The sixth term is obtained by setting $j = 5$. It is

$$\binom{12}{5}x^{12-5}(2y)^5 = 792x^7(2y)^5$$

Example $\sum_{j=0}^{14}\binom{n}{j} = (1+1)^{14} = 2^{14}$.

If n is not a positive integer, the sum formula no longer applies and an infinite series results for $(a + b)^n$. The coefficients are obtained from the first formulas in this case.

Example $(1 + x)^{1/2} = 1 + \frac{1}{2}x - \frac{1}{2} \cdot \frac{1}{4}x^2 + \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{3}{6}x^3 \cdots$ (convergent for $x^2 < 1$).

Additional discussion is under "Infinite Series."

PROGRESSIONS

An arithmetic progression is a succession of terms such that each term, except the first, is derivable from the preceding by the addition of a quantity d called the common difference. All arithmetic progressions have the form a, $a + d$, $a + 2d$, $a + 3d$, ... With a = first term, l = last term, d = common difference, n = number of terms, and s = sum of the terms, the following relations hold:

$$l = a + (n-1)d = -\frac{d}{2} + \sqrt{2ds + \left(a - \frac{d}{2}\right)^2} = \frac{s}{n} + \frac{(n-1)}{2}d$$

$$s = \frac{n}{2}[2a + (n-1)d] = \frac{n}{2}(a + l) = \frac{n}{2}[2l - (n-1)d]$$

$$a = l - (n-1)d = \frac{s}{n} - \frac{(n-1)d}{2} = \frac{2s}{n} - l$$

$$d = \frac{l-a}{n-1} = \frac{2(s-an)}{n(n-1)} = \frac{2(nl-s)}{n(n-1)}$$

$$n = \frac{l-a}{d} + 1 = \frac{2s}{l+a} = \frac{2l + d + \sqrt{(2l+d)^2 - 8ds}}{2d}$$

The **arithmetic mean or average** of two numbers a, b is $(a + b)/2$; of n numbers $a_1, \ldots, a_n$ is $(a_1 + a_2 + \cdots + a_n)/n$.

A **geometric progression** is a succession of terms such that each term, except the first, is derivable from the preceding by the multiplication of a quantity r called the common ratio. All such progressions have the form a, ar, $ar^2, \ldots, ar^{n-1}$. With a = first term, l = last term, r = ratio, n = number of terms, s = sum of the terms, the following relations hold:

$$l = ar^{n-1} = \frac{[a + (r-1)s]}{r} = \frac{(r-1)sr^{n-1}}{r^n - 1}$$

$$s = \frac{a(r^n - 1)}{r - 1} = \frac{a(1 - r^n)}{1 - r} = \frac{rl - a}{r - 1} = \frac{lr^n - l}{r^n - r^{n-1}}$$

$$a = \frac{l}{r^{n-l}} = \frac{(r-1)s}{r^n - 1} \qquad r = \frac{s-a}{s-l} \qquad \log r = \frac{\log l - \log a}{n-1}$$

$$n = \frac{\log l - \log a}{\log r} + 1 = \frac{\log[a + (r-1)s] - \log a}{\log r}$$

The geometric mean of two nonnegative numbers a, b is $\sqrt{ab}$; of n numbers is $(a_1a_2 \ldots a_n)^{1/n}$.

Example Find the sum of $1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{64}$. Here $a = 1$, $r = \frac{1}{2}$, $n = 7$. Thus

$$s = \frac{\frac{1}{2}(\frac{1}{64}) - 1}{\frac{1}{2} - 1} = 127/64$$

$$s = a + ar + ar^2 + \cdots + ar^{n-1} = \frac{a}{1-r} - \frac{ar^n}{1-r}$$

If $|r| < 1$, then $\lim_{n\to\infty} s = \frac{a}{1-r}$

which is called the sum of the infinite geometric progression.

Example The present worth (PW) of a series of cash flows C_k at the end of year k is

$$\text{PW} = \sum_{k=1}^{n}\frac{C_k}{(1+i)^k}$$

where i is an assumed interest rate. (Thus the present worth always requires specification of an interest rate.) If all the payments are the same, $C_k = R$, the present worth is

$$\text{PW} = R\sum_{k=1}^{n}\frac{1}{(1+i)^k}$$

This can be rewritten as

$$\text{PW} = \frac{R}{1+i}\sum_{k=1}^{n}\frac{1}{(1+i)^{k-1}} = \frac{R}{1+i}\sum_{j=0}^{n-1}\frac{1}{(1+i)^j}$$

This is a geometric series with $r = 1/(1 + i)$ and $a = R/(1 + i)$. The formulas above give

$$\text{PW } (=s) = \frac{R}{i}\frac{(1+i)^n - 1}{(1+i)^n}$$

The same formula applies to the value of an annuity (PW) now, to provide for equal payments R at the end of each of n years, with interest rate i.

A progression of the form a, $(a + d)r$, $(a + 2d)r^2$, $(a + 3d)r^3$, etc., is a combined arithmetic and geometric progression. The sum of n such terms is

$$s = \frac{a - [a + (n-1)d]r^n}{I - r} + \frac{rd(1 - r^{n-1})}{(1-r)^2}$$

If $|r| < 1$, $\lim_{n\to\infty} s = \frac{a}{1-r} + rd/(1-r)^2$.

The non-zero numbers $a, b, c,$ etc., form a harmonic progression if their reciprocals $1/a, 1/b, 1/c,$ etc., form an arithmetic progression.

Example The progression $1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \ldots, \frac{1}{31}$ is harmonic since 1, 3, 5, 7, . . . , 31 form an arithmetic progression.

The **harmonic mean** of two numbers a, b is $2ab/(a+b)$.

PERMUTATIONS, COMBINATIONS, AND PROBABILITY

Each separate arrangement of all or a part of a set of things is called a **permutation.** The number of permutations of n things taken r at a time, written

$$P(n, r) = \frac{n!}{(n-r)!} = n(n-1)(n-2)\cdots(n-r+1)$$

Example The permutations of a, b, c two at a time are $ab, ac, ba, ca, cb,$ and $bc.$ The formula is $P(3,2) = 3!/1! = 6.$ The permutations of a, b, c three at a time are $abc, bac, cab, acb, bca,$ and $cba.$

Each separate selection of objects that is possible irrespective of the order in which they are arranged is called a combination. The number of combinations of n things taken r at a time, written $C(n, r) = n!/[r!(n-r)!]$.

Example The combinations of a, b, c taken 2 at a time are $ab, ac, bc;$ taken 3 at a time is $abc.$

An important relation is $r!\, C(n, r) = P(n, r)$.

If an event can occur in p ways and fail to occur in q ways, all ways being equally likely, the *probability* of its occurrence is $p/(p+q)$, and that of its failure $q/(p+q)$.

Example Two dice may be thrown in 36 separate ways. What is the probability of throwing such that their sum is 7? Seven may arise in 6 ways: 1 and 6, 2 and 5, 3 and 4, 4 and 3, 5 and 2, 6 and 1. The probability of shooting 7 is $\frac{1}{6}$.

THEORY OF EQUATIONS

Linear Equations A linear equation is one of the first degree (i.e., only the first powers of the variables are involved), and the process of obtaining definite values for the unknown is called solving the equation. Every linear equation in one variable is written $Ax + B = 0$ or $x = -B/A.$ Linear equations in n variables have the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

The solution of the system may then be found by elimination or matrix methods if a solution exists (see "Matrix Algebra and Matrix Computations").

Quadratic Equations Every quadratic equation in one variable is expressible in the form $ax^2 + bx + c = 0.\ a \neq 0.$ This equation has two solutions, say, $x_1, x_2,$ given by

$$\left.\begin{matrix} x_1 \\ x_2 \end{matrix}\right\} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

If a, b, c are real, the discriminant $b^2 - 4ac$ gives the character of the roots. If $b^2 - 4ac > 0$, the roots are real and unequal. If $b^2 - 4ac < 0$, the roots are complex conjugates. If $b^2 - 4ac = 0$ the roots are **real** and **equal.**

Two quadratic equations in two variables can in general be solved only by numerical methods (see "Numerical Analysis and Approximate Methods"). If one equation is of the first degree, the other of the second degree, a solution may be obtained by solving the first for one unknown. This result is substituted in the second equation and the resulting quadratic equation solved.

Cubic Equations A cubic equation, in one variable, has the form $x^3 + bx^2 + cx + d = 0.$ Every cubic equation having complex coefficients has three complex roots. If the coefficients are real numbers, then at least one of the roots must be real. The cubic equation $x^3 + bx^2 + cx + d = 0$ may be reduced by the substitution $x = y - (b/3)$ to the form $y^3 + py + q = 0,$ where $p = \frac{1}{3}(3c - b^2),\ q = \frac{1}{27}(27d - 9bc + 2b^3).$ This equation has the solutions $y_1 = A + B,\ y_2 = -\frac{1}{2}(A + B) + (i\sqrt{3}/2)(A - B),$ $y_3 = -\frac{1}{2}(A + B) - (i\sqrt{3}/2)(A - B),$ where $i^2 = -1,\ A = \sqrt[3]{-q/2 + \sqrt{R}},$ $B = \sqrt[3]{-q/2 - \sqrt{R}},$ and $R = (p/3)^3 + (q/2)^2.$ If b, c, d are all real and if $R > 0,$ there are one real root and two conjugate complex roots; if $R = 0,$ there are three real roots, of which at least two are equal; if $R < 0,$ there are three real unequal roots. If $R < 0,$ these formulas are impractical. In this case, the roots are given by $y_k = \mp\ 2\sqrt{-p/3}\cos[(\phi/3) + 120k],\ k = 0, 1, 2$ where

$$\phi = \cos^{-1}\sqrt{\frac{q^2/4}{-p^3/27}}$$

and the upper sign applies if $q > 0,$ the lower if $q < 0.$

Example $x^3 + 3x^2 + 9x + 9 = 0$ reduces to $y^3 + 6y + 2 = 0$ under $x = y - 1.$ Here $p = 6,\ q = 2,\ R = 9.$ Hence $A = \sqrt[3]{2},\ B = \sqrt[3]{-4}.$ The desired roots in y are $\sqrt[3]{2} - \sqrt[3]{-4}$ and $-\frac{1}{2}(\sqrt[3]{2} - \sqrt[3]{4}) \pm (i\sqrt{3}/2)(\sqrt[3]{2} + \sqrt[3]{4}).$ The roots in x are $x = y - 1.$

Example $y^3 - 7y + 7 = 0.\ p = -7,\ q = 7,\ R < 0.$ Hence

$$x_k = -\sqrt{\frac{28}{3}}\cos\left(\frac{\phi}{3} + 120k\right)$$

where

$$\phi = \sqrt{\frac{27}{28}},\ \frac{\phi}{3} = 3°37'52''.$$

The roots are approximately $-3.048916,\ 1.692020,$ and $1.356897.$

Example Many equations of state involve solving cubic equations for the compressibility factor $Z.$ For example, the Redlich-Kwong-Soave equation of state requires solving

$$Z^3 - Z^2 + cZ + d = 0, \quad d < 0$$

where c and d depend on critical constants of the chemical species. In this case, only positive solutions, $Z > 0,$ are desired.

Quartic Equations See Ref. 118.

General Polynomials of the *n*th Degree Denote the general polynomial equation of degree n by

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$$

If $n > 4,$ there is no formula which gives the roots of the general equation. For fourth and higher order (even third order), the roots can be found numerically (see "Numerical Analysis and Approximate Methods"). However, there are some general theorems that may prove useful.

Remainder Theorems When $P(x)$ is a polynomial and $P(x)$ is divided by $x - a$ until a remainder independent of x is obtained, this remainder is equal to $P(a).$

Example $P(x) = 2x^4 - 3x^2 + 7x - 2$ when divided by $x + 1$ (here $a = -1$) results in $P(x) = (x + 1)(2x^3 - 2x^2 - x + 8) - 10$ where -10 is the remainder. It is easy to see that $P(-1) = -10.$

Factor Theorem If $P(a)$ is zero, the polynomial $P(x)$ has the factor $x - a.$ In other words, if a is a root of $P(x) = 0,$ then $x - a$ is a factor of $P(x).$

If a number a is found to be a root of $P(x) = 0,$ the division of $P(x)$ by $(x - a)$ leaves a polynomial of degree one less than that of the original equation, i.e., $P(x) = Q(x)(x - a).$ Roots of $Q(x) = 0$ are clearly roots of $P(x) = 0.$

Example $P(x) = x^3 - 6x^2 + 11x - 6 = 0$ has the root + 3. Then $P(x) = (x - 3)(x^2 - 3x + 2).$ The roots of $x^2 - 3x + 2 = 0$ are 1 and 2. The roots of $P(x)$ are therefore 1, 2, 3.

Fundamental Theorem of Algebra Every polynomial of degree n has exactly n real or complex roots, counting multiplicities.

Every polynomial equation $a_0x^n + a_1x^{n-1} + \cdots + a_n = 0$ with *rational coefficients* may be rewritten as a polynomial, of the same degree, with *integral coefficients* by multiplying each coefficient by the least common multiple of the denominators of the coefficients.

Example The coefficients of $\frac{1}{2}x^4 + \frac{1}{3}x^3 - \frac{5}{6}x^2 + 2x - \frac{1}{6} = 0$ are rational numbers. The least common multiple of the denominators is $2 \times 3 = 6$. Therefore, the equation is equivalent to $9x^4 + 14x^3 - 5x^2 + 12x - 1 = 0$.

Upper Bound for the Real Roots Any number that exceeds all the roots is called an upper bound to the real roots. If the coefficients of a polynomial equation are all of like sign, there is no positive root. Such equations are excluded here since zero is the upper bound to the real roots. If the coefficient of the highest power of $P(x) = 0$ is negative, replace the equation by $-P(x) = 0$.

If in a polynomial $P(x) = c_0x^n + c_1x^{n-1} + \cdots + c_{n-1}x + c_n = 0$, with $c_0 > 0$, the first negative coefficient is preceded by k coefficients which are positive or zero, and if G denotes the greatest of the numerical values of the negative coefficients, then each real root is less than $1 + \sqrt[k]{G/c_0}$.

A lower bound to the negative roots of $P(x) = 0$ may be found by applying the rule to $P(-x) = 0$.

Example $P(x) = x^7 + 2x^5 + 4x^4 - 8x^2 - 32 = 0$. Here $k = 5$ (since 2 coefficients are zero), $G = 32$, $c_0 = 1$. The upper bound is $1 + \sqrt[5]{32} = 3$. $P(-x) = -x^7 - 2x^5 + 4x^4 - 8x^2 - 32 = 0$. $-P(-x) = x^7 + 2x^5 - 4x^4 + 8x^2 + 32 = 0$. Here $k = 3$, $G = 4$, $c_0 = 1$. The lower bound is $-(1 + \sqrt[3]{4}) \approx -2.587$. Thus all real roots r lie in the range $-2.587 < r < 3$.

Descartes Rule of Signs The number of positive real roots of a polynomial equation with real coefficients either is equal to the number v of its variations in sign or is less than v by a positive even integer. The number of negative roots of $P(x) = 0$ either is equal to the number of variations of sign of $P(-x)$ or is less than that number by a positive even integer.

Example $P(x) = x^4 + 3x^3 + x - 1 = 0$. $v = 1$; so $P(x)$ has one positive root. $P(-x) = x^4 - 3x^3 - x - 1$. Here $v = 1$; so $P(x)$ has one negative root. The other two roots are complex conjugates.

Example $P(x) = x^4 - x^2 + 10x - 4 = 0$. $v = 3$; so $P(x)$ has three or one positive roots. $P(-x) = x^4 - x^2 - 10x - 4$. $v = 1$; so $P(x)$ has exactly one negative root.

Numerical methods are often used to find the roots of polynomials. A detailed discussion of these techniques is given under "Numerical Analysis and Approximate Methods."

Determinants Consider the system of two linear equations

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2$$

If the first equation is multiplied by a_{22} and the second by $-a_{12}$ and the results added, we obtain

$$(a_{11}a_{22} - a_{21}a_{12})x_1 = b_1a_{22} - b_2a_{12}$$

The expression $a_{11}a_{22} - a_{21}a_{12}$ may be represented by the symbol

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

This symbol is called a determinant of second order. The value of the square array of n^2 quantities a_{ij}, where $i = 1, \ldots, n$ is the row index, $j = 1, \ldots, n$ the column index, written in the form

$$|A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \cdots a_{1n} \\ a_{21} & a_{22} & \cdots\cdots\ a_{2n} \\ \vdots & & \\ a_{n1} & a_{n2} & a_{n3} \cdots a_{nn} \end{vmatrix}$$

is called a determinant. The n^2 quantities a_{ij} are called the elements of the determinant. In the determinant $|A|$ let the ith row and jth column be deleted and a new determinant be formed having $n - 1$ rows and columns. This new determinant is called the minor of a_{ij} denoted M_{ij}.

Example $\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$ The minor of a_{23} is $M_{23} = \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix}$

The cofactor A_{ij} of the element a_{ij} is the signed minor of a_{ij} determined by the rule $A_{ij} = (-1)^{i+j}M_{ij}$. The *value* of $|A|$ is obtained by forming any of the equivalent expressions $\sum_{j=1}^{n} a_{ij}A_{ij}$, $\sum_{i=1}^{n} a_{ij}A_{ij}$, where the elements a_{ij} must be taken from a single row or a single column of A.

Example

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{31}A_{31} + a_{32}A_{32} + a_{33}A_{33}$$
$$= a_{31}\begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} - a_{32}\begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} + a_{33}\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

In general, A_{ij} will be determinants of order $n - 1$, but they may in turn be expanded by the rule. Also,

$$\sum_{j=1}^{n} a_{ji}A_{jk} = \sum_{j=1}^{n} a_{ij}A_{jk} = \begin{cases} |A| & i = k \\ 0 & i \neq k \end{cases}$$

Fundamental Properties of Determinants

1. The value of a determinant $|A|$ is not changed if the rows and columns are interchanged.
2. If the elements of one row (or one column) of a determinant are all zero, the value of $|A|$ is zero.
3. If the elements of one row (or column) of a determinant are multiplied by the same constant factor, the value of the determinant is multiplied by this factor.
4. If one determinant is obtained from another by interchanging any two rows (or columns), the value of either is the negative of the value of the other.
5. If two rows (or columns) of a determinant are identical, the value of the determinant is zero.
6. If two determinants are identical except for one row (or column), the sum of their values is given by a single determinant obtained by adding corresponding elements of dissimilar rows (or columns) and leaving unchanged the remaining elements.

Example

$$\begin{vmatrix} 3 & 2 \\ 1 & 5 \end{vmatrix} + \begin{vmatrix} 4 & 2 \\ 7 & 5 \end{vmatrix} = 13 + 6 = 19 \quad \text{Directly}$$

$$\begin{vmatrix} 7 & 2 \\ 8 & 5 \end{vmatrix} = 35 - 16 = 19 \quad \text{By rule 6}$$

7. The value of a determinant is not changed if to the elements of any row (or column) are added a constant multiple of the corresponding elements of any other row (or column).
8. If all elements but one in a row (or column) are zero, the value of the determinant is the product of that element times its cofactor.

The evaluation of determinants using the definition is quite laborious. The labor can be reduced by applying the fundamental properties just outlined.

The solution of n linear equations (not all b_i zero)

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$
$$\vdots\vdots \qquad\qquad \vdots$$
$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n$$

where $|A| = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \\ a_{n_1} & \cdots & a_{nn} \end{vmatrix} \neq 0$

has a unique solution given by $x_1 = |B_1|/|A|$, $x_2 = |B_2|/|A|$, $\ldots$, $x_n = |B_n|/|A|$, where B_k is the determinant obtained from A by replacing its kth column by $b_1, b_2, \ldots, b_n$. This technique is called **Cramer's rule.** It requires more labor than the method of elimination and should not be used for computations.

ANALYTIC GEOMETRY

REFERENCES: 108, 188, 193, 260, 261, 268, 274, 282.

Analytic geometry uses algebraic equations and methods to study geometric problems. It also permits one to visualize algebraic equations in terms of geometric curves, which frequently clarifies abstract concepts.

PLANE ANALYTIC GEOMETRY

Coordinate Systems The basic concept of analytic geometry is the establishment of a one-to-one correspondence between the points of the plane and number pairs (x, y). This correspondence may be done in a number of ways. The rectangular or cartesian coordinate system consists of two straight lines intersecting at right angles (Fig. 3-12). A point is designated by (x, y), where x (the abscissa) is the distance of the point from the y axis measured parallel to the x axis, positive if to the right, negative to the left. y (ordinate) is the distance of the point from the x axis, measured parallel to the y axis, positive if above, negative if below the x axis. The **quadrants** are labeled 1, 2, 3, 4 in the drawing, the coordinates of points in the various quadrants having the depicted signs. Another common coordinate system is the polar coordinate system (Fig. 3-13). In this system the position of a point is designated by the pair (r, θ), $r = \sqrt{x^2 + y^2}$ being the distance to the origin 0(0,0) and θ being the angle the line r makes with the positive x axis (polar axis). To change from polar to rectangular coordinates, use $x = r \cos \theta$ and $y = r \sin \theta$. To change from rectangular to polar coordinates, use $r = \sqrt{x^2 + y^2}$ and $\theta = \tan^{-1}(y/x)$ if $x \neq 0$; $\theta = \pi/2$ if $x = 0$. The distance between two points (x_1, y_1), (x_2, y_2) is defined by $d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ in rectangular coordinates or by $d = \sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos(\theta_1 - \theta_2)}$ in polar coordinates. Other coordinate systems are sometimes used. For example, on the surface of a sphere latitude and longitude prove useful.

The Straight Line (Fig. 3-14) The slope m of a straight line is the tangent of the inclination angle θ made with the positive x axis. If (x_1, y_1) and (x_2, y_2) are any two points on the line, slope $= m = (y_2 - y_1)/(x_2 - x_1)$. The slope of a line parallel to the x axis is zero; parallel to the y axis, it is undefined. Two lines are parallel if and only if they have the same slope. Two lines are perpendicular if and only if the product of their slopes is -1 (the exception being that case when the lines are parallel to the coordinate axes). Every equation of the type $Ax + By + C = 0$ represents a straight line, and every straight line has an equation of this form. A straight line is determined by a variety of conditions:

	Given conditions	Equation of line
(1)	Parallel to x axis	$y =$ constant
(2)	Parallel y axis	$x =$ constant
(3)	Point (x_1, y_1) and slope m	$y - y_1 = m(x - x_1)$
(4)	Intercept on y axis $(0, b)$, m	$y = mx + b$
(5)	Intercept on x axis $(a, 0)$, m	$y = m(x - a)$
(6)	Two points (x_1, y_1), (x_2, y_2)	$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$
(7)	Two intercepts $(a, 0)$, $(0, b)$	$x/a + y/b = 1$

The angle β a line with slope m_1 makes with a line having slope m_2 is given by $\tan \beta = (m_2 - m_1)/(m_1 m_2 + 1)$. A line is determined if the length and direction of the perpendicular to it (the normal) from the origin are given (see Fig. 3-15). Let p = length of the perpendicular and α the angle that the perpendicular makes with the positive x axis. The equation of the line is $x \cos \alpha + y \sin \alpha = p$. The equation of a line perpendicular to a given line of slope m and passing through a point (x_1, y_1) is $y - y_1 = -(1/m)(x - x_1)$. The distance from a point (x_1, y_1) to a line with equation $Ax + by + C = 0$ is

$$d = \frac{|Ax_1 + By_1 + C|}{\sqrt{A^2 + B^2}}$$

Example If it is known that centigrade C and Fahrenheit F are linearly related and when C = 0°, F = 32°; C = 100°, F = 212°, find the equation relating C and F and that point where C = F. By using the two-point form, the equation is

$$F - 32 = \frac{212 - 32}{100 - 0}(C - 0)$$

or F = ⅘C + 32. Equivalently

$$C - 0 = \frac{100 - 0}{212 - 32}(F - 32)$$

or C = ⅝(F − 32). Letting C = F, we have from either equation F = C = −40.

Occasionally some nonlinear algebraic equations can be reduced to linear equations under suitable substitutions or changes of variables. In other words, certain curves become the graphs of lines if the scales or coordinate axes are appropriately transformed.

Example Consider $y = bx^n$. $B = \log b$. Taking logarithms $\log y = n \log x + \log b$. Let $Y = \log y$, $X = \log x$, $B = \log b$. The equation then has the form $Y = nX + B$, which is a linear equation. Consider $k = k_0 \exp(-E/RT)$, taking logarithms $\log_e k = \log_e k_0 - E/(RT)$. Let $Y = \log_e k$, $B = \log_e k_0$, and $m = -E/R$, $X = 1/T$, and the result is $Y = mX + B$. Next consider $y = a + bx^n$. If the substitution $t = x^n$ is made, then the graph of y is a straight line versus t.

Asymptotes The limiting position of the tangent to a curve as the point of contact tends to an infinite distance from the origin is called an **asymptote.** If the equation of a given curve can be expanded in a Laurent power series such that

$$f(x) = \sum_{k=0}^{n} a_k x^k + \sum_{k=0}^{n} \frac{b_k}{x^k}$$

and

$$\lim_{x \to \infty} f(x) = \sum_{k=0}^{n} a_k x^k$$

then the equation of the asymptote is $y = \sum_{k=0}^{n} a_k x^k$. If $n = 1$, then the asymptote is (in general oblique) a line. In this case, the equation of the asymptote may be written as

$$y = mx + b \qquad m = \lim_{x \to \infty} f'(x)$$

$$b = \lim_{x \to \infty} [f(x) - x f'(x)]$$

Geometric Properties of a Curve When the Equation Is Given The analysis of the properties of an equation is facilitated by the investigation of the equation by using the following techniques:

1. *Points of maximum, minimum, and inflection.* These may be investigated by means of the calculus.

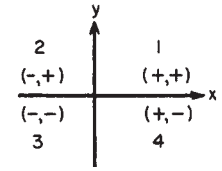

FIG. 3-12 Rectangular coordinates.

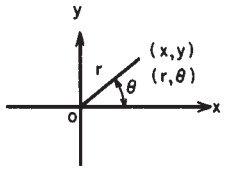

FIG. 3-13 Polar coordinates.

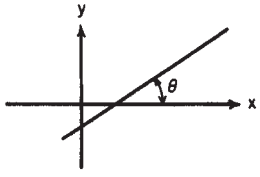

FIG. 3-14 Straight line.

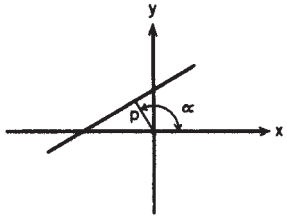

FIG. 3-15 Determination of line.

2. *Symmetry.* Let $F(x, y) = 0$ be the equation of the curve.

Condition on $F(x, y)$	Symmetry
$F(x, y) = F(-x, y)$	With respect to y axis
$F(x, y) = F(x, -y)$	With respect to x axis
$F(x, y) = F(-x, -y)$	With respect to origin
$F(x, y) = F(y, x)$	With respect to the line $y = x$

3. *Extent.* Only real values of x and y are considered in obtaining the points (x, y) whose coordinates satisfy the equation. The extent of them may be limited by the condition that negative numbers do not have real square roots.
4. *Intercepts.* Find those points where the curves of the function cross the coordinate axes.
5. *Asymptotes.* See preceding discussion.
6. *Direction at a point.* This may be found from the derivative of the function at a point. This concept is useful for distinguishing among a family of similar curves.

Example $y^2 = (x^2 + 1)/(x^2 - 1)$ is symmetric with respect to the x and y axis, the origin, and the line $y = x$. It has the vertical asymptotes $x = \pm 1$. When $x = 0$, $y^2 = -1$; so there are no y intercepts. If $y = 0$, $(x^2 + 1)/(x^2 - 1) = 0$; so there are no x intercepts. If $|x| < 1$, y^2 is negative; so $|x| > 1$. From $x^2 = (y^2 + 1)/(y^2 - 1)$, $y = \pm 1$ are horizontal asymptotes and $|y| > 1$. As $x \to 1^+$, $y \to +\infty$; as $x \to +\infty$, $y \to +1$. The graph is given in Fig. 3-16.

Conic Sections The curves included in this group are obtained from plane sections of the cone. They include the circle, ellipse, parabola, hyperbola, and degeneratively the point and straight line. A **conic** is the locus of a point whose distance from a fixed point called the **focus** is in a constant ratio to its distance from a fixed line, called the directrix. This ratio is the eccentricity e. If $e = 0$, the conic is a circle; if $0 < e < 1$, the conic is an ellipse; if $e = 1$, the conic is a parabola; if $e > 1$, the conic is a hyperbola. Every conic section is representable by an equation of second degree. Conversely, every equation of second degree in two variables represents a conic. The general equation of the second degree is $Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$. Let Δ be defined as the determinant

$$\Delta = \begin{vmatrix} 2A & B & D \\ B & 2C & E \\ D & E & 2F \end{vmatrix}$$

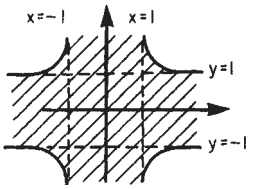

FIG. 3-16 Graph of $y^2 = (x^2 + 1)/(x^2 - 1)$

The table characterizes the curve represented by the equation.

	$B^2 - 4AC < 0$	$B^2 - 4AC = 0$	$B^2 - 4AC > 0$
$\Delta \neq 0$	$A\Delta < 0$ $A \neq C$, an ellipse $A\Delta < 0$ $A = C$, a circle $A\Delta > 0$, no locus	Parabola	Hyperbola
$\Delta = 0$	Point	2 parallel lines if $Q = D^2 + E^2 - 4(A + C)F > 0$ 1 straight line if $Q = 0$, no locus if $Q < 0$	2 intersecting straight lines

Example $3x^2 + 4xy - 2y^2 + 3x - 2y + 7 = 0$.

$$\Delta = \begin{vmatrix} 6 & 4 & 3 \\ 4 & -4 & -2 \\ 3 & -2 & 14 \end{vmatrix} = -596 \neq 0,\ B^2 - 4AC = 40 > 0$$

The curve is therefore a hyperbola.

To translate the axes to a new origin at (h, k), substitute for x and y in the original equation $x + h$ and $y + k$. Translation of the axes can always be accomplished to eliminate the linear terms in the second-degree equation in two variables having no xy term.

Example $x^2 + y^2 + 2x - 4y + 2 = 0$. Rewrite this as $x^2 + 2x + 1 + y^2 - 4y + 4 - 5 + 2 = 0$ or $(x + 1)^2 + (y - 2)^2 = 3$. Let $u = x + 1$, $v = y - 2$. Then $u^2 + v^2 = 3$. The axis has been translated to the new origin $(-1, 2)$.

The type of curve determined by a specific equation of the second degree can also be easily determined by reducing it to a standard form by translation and/or rotation. In the case in which the equation has no xy term, the procedure is merely to complete the squares of the terms in x and y separately.

To rotate the axes through an angle α, substitute for x the quantity $x \cos \alpha - y \sin \alpha$ and for y the quantity $x \sin \alpha + y \cos \alpha$. A rotation of the axes through $\alpha = \frac{1}{2} \cot^{-1} (A - C)/B$ will eliminate the cross-product term in the general second-degree equation.

Example Consider $3x^2 + 2xy + y^2 - 2x + 3y = 7$. A rotation of axes through $\alpha = \frac{1}{2} \cot^{-1} 1 = 22\frac{1}{2}°$ eliminates the cross-product term.

The following tabulation gives the form of the more common equations.

Polar equation	Type of curve
(1) $r = a$	Circle
(2) $r = 2a \cos \theta$	Circle
(3) $r = 2a \sin \theta$	Circle
(4) $r^2 - 2br \cos (\theta - \beta) + b^2 - a^2 = 0$	Circle at (b, β), radius a
(5) $r = \dfrac{ke}{1 - e \cos \theta}$	$e = 1$ parabola $0 < e < 1$ ellipse $e > 1$ hyperbola

Some common equations in parametric form are given below.

(1) $(x - h)^2 + (y - k)^2 = a^2$	$x = h + a \cos \theta$ $y = k + a \sin \theta$	Circle (Fig. 3-23) Parameter is angle θ.
(2) $\dfrac{(x - h)^2}{a^2} + \dfrac{(y - k)^2}{b^2} = 1$	$x = h + a \cos \phi$ $y = k + a \sin \phi$	Ellipse (Fig. 3-20) Parameter is angle ϕ.
	$x = \dfrac{-at}{\sqrt{t^2 + 1}}$	Circle Parameter is $t = \dfrac{dy}{dx}$ = slope of tangent at (x, y).
(3) $z^2 + y^2 = a^2$	$y = \dfrac{a}{\sqrt{t^2 + 1}}$	
(4) $y = a \cosh \dfrac{x}{a}$	$x = a \sinh^{-1} \dfrac{s}{a}$ $y^2 = a^2 + s^2$	Catenary (Fig. 3-24; such as hanging cable under gravity) Parameter s = arc length from $(0, a)$ to (x, y). See Fig. 3-24.
(5) Cycloid	$x = a(\phi - \sin \phi)$ $y = a(1 - \cos \phi)$	

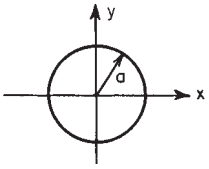

FIG. 3-17 Circle center (0,0) $r = a$.

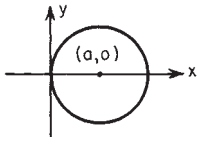

FIG. 3-18 Circle center $(a,0)$ $r = 2a \cos \theta$.

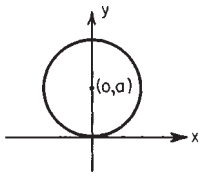

FIG. 3-19 Circle center $(0,a)$ $r = 2a \sin \theta$.

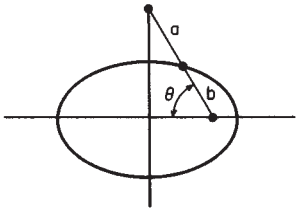

FIG. 3-20 Ellipse, $0 < e < 1$.

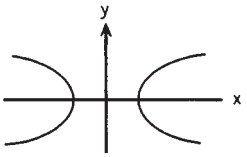

FIG. 3-21 Hyperbola, $e > 1$, $r = ke/(1 - e \cos \theta)$.

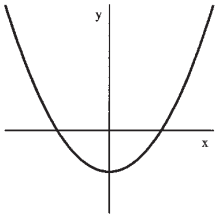

FIG. 3-22 Parabola, $e = 1$.

Circle at (b, β), radius a: $r^2 - 2br \cos(\theta - \beta) + b^2 - a^2 = 0$.

Graphs of Polar Equations The equation $r = 0$ corresponds to $x = 0$, $y = 0$ regardless of θ. The same point may be represented in several different ways; thus the point $(2, \pi/3)$ or $(2, 60°)$ has the following representations: $(2, 60°)$, $(2, -300°)$. These are summarized in $(2, 60° + n\ 360°)$, $n = 0, \pm1, \pm2$, or in radian measure $[2, (\pi/3) + 2n\pi]$, $n = 0, \pm1, \pm2$. Plotting of polar equations can be facilitated by the following steps:

1. Find those points where r is a maximum or minimum.
2. Find those values of θ where $r = 0$, if any.
3. Symmetry: The curve is symmetric about the origin if the equation is unchanged when θ is replaced by $\theta \pm \pi$, symmetric about the x axis if the equation is unchanged when θ is replaced by $-\theta$, and symmetric about the y axis if the equation is unchanged when θ is replaced by $\pi - \theta$.

Parametric Equations It is frequently useful to write the equations of a curve in terms of an auxiliary variable called a parameter. For example, a circle of radius a, center at $(0, 0)$, can be written in the equivalent form $x = a \cos \theta$, $y = a \sin \phi$ where θ is the parameter. Similarly, $x = a \cos \phi$, $y = b \sin \phi$ are the parametric equations of the ellipse $x^2/a^2 + y^2/b^2 = 1$ with parameter ϕ.

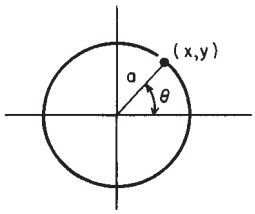

FIG. 3-23 Circle.

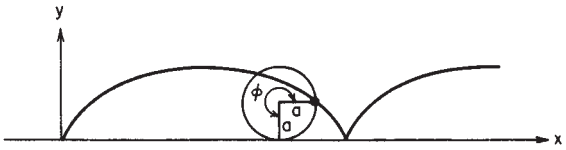

FIG. 3-24 Cycloid.

SOLID ANALYTIC GEOMETRY

Coordinate Systems The commonly used coordinate systems are three in number. Others may be used in specific problems (see Ref. 212). The **rectangular** (cartesian) system (Fig. 3-25) consists of mutually orthogonal axes x, y, z. A triple of numbers (x, y, z) is used to represent each point. The **cylindrical** coordinate system $(r, \theta, z;$ Fig. 3-26) is frequently used to locate a point in space. These are essentially the polar coordinates (r, θ) coupled with the z coordinate. As

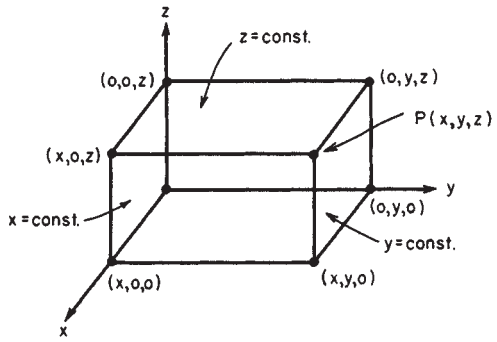

FIG. 3-25 Cartesian coordinates.

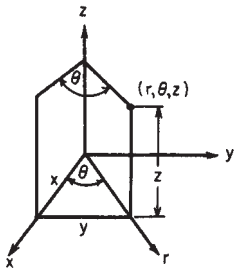

FIG. 3-26 Cylindrical coordinates.

before, $x = r\cos\theta$, $y = r\sin\theta$, $z = z$ and $r^2 = x^2 + y^2$, $y/x = \tan\theta$. If r is held constant and θ and z are allowed to vary, the locus of (r, θ, z) is a right circular cylinder of radius r along the z axis. The locus of $r = C$ is a circle, and θ = constant is a plane containing the z axis and making an angle θ with the xz plane. Cylindrical coordinates are convenient to use when the problem has an axis of symmetry.

The **spherical** coordinate system is convenient if there is a point of symmetry in the system. This point is taken as the origin and the coordinates (ρ, ϕ, θ) illustrated in Fig. 3-27. The relations are $x = \rho\sin\phi\cos\theta$, $y = \rho\sin\phi\sin\theta$, $z = \rho\cos\phi$, and $r = \rho\sin\phi$. θ = constant is a plane containing the z axis and making an angle θ with the xz plane. ϕ = constant is a cone with vertex at 0. ρ = constant is the surface of a sphere of radius ρ, center at the origin 0. Every point in the space may be given spherical coordinates restricted to the ranges $0 \le \phi \le \pi$, $\rho \ge 0$, $0 \le \theta < 2\pi$.

Lines and Planes The distance between two points (x_1, y_1, z_1), (x_2, y_2, z_2) is $d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$. There is nothing in the geometry of three dimensions quite analogous to the slope of a line in the plane case. Instead of specifying the direction of a line by a trigonometric function evaluated for one angle, a trigonometric function evaluated for three angles is used. The angles α, β, γ that a line segment makes with the positive x, y, and z axes, respectively, are called the **direction angles** of the line, and $\cos\alpha$, $\cos\beta$, $\cos\gamma$ are called the **direction cosines.** Let (x_1, y_1, z_1), (x_2, y_2, z_2) be on the line. Then $\cos\alpha = (x_2 - x_1)/d$, $\cos\beta = (y_2 - y_1)/d$, $\cos\gamma = (z_2 - z_1)/d$, where d = the distance between the two points. Clearly $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1$. If two lines are specified by the direction cosines $(\cos\alpha_1, \cos\beta_1, \cos\gamma_1)$, $(\cos\alpha_2, \cos\beta_2, \cos\gamma_2)$, then the angle θ between the lines is $\cos\theta = \cos\alpha_1\cos\alpha_2 + \cos\beta_1\cos\beta_2 + \cos\gamma_1\cos\gamma_2$. Thus the lines are perpendicular if and only if $\theta = 90°$ or $\cos\alpha_1\cos\alpha_2 + \cos\beta_1\cos\beta_2 + \cos\gamma_1\cos\gamma_2 = 0$. The equation of a line with direction cosines $(\cos\alpha, \cos\beta, \cos\gamma)$ passing through (x_1, y_1, z_1) is $(x - x_1)/\cos\alpha = (y - y_1)/\cos\beta = (z - z_1)/\cos\gamma$.

The equation of every plane is of the form $Ax + By + Cz + D = 0$. The numbers

$$\frac{A}{\sqrt{A^2 + B^2 + C^2}}, \frac{B}{\sqrt{A^2 + B^2 + C^2}}, \frac{C}{\sqrt{A^2 + B^2 + C^2}}$$

are direction cosines of the normal lines to the plane. The plane through the point (x_1, y_1, z_1) whose normals have these as direction cosines is $A(x - x_1) + B(y - y_1) + C(z - z_1) = 0$.

Example Find the equation of the plane through $(1, 5, -2)$ perpendicular to the line $(x + 9)/7 = (y - 3)/-1 = z/8$. The numbers $(7, -1, 8)$ are called **direction numbers.** They are a constant multiple of the direction cosines. $\cos\alpha = 7/114$, $\cos\beta = -1/114$, $\cos\gamma = 8/114$. The plane has the equation $7(x - 1) - 1(y - 5) + 8(z + 2) = 0$ or $7x - y + 8z + 14 = 0$.

The distance from the point (x_1, y_1, z_1) to the plane $Ax + By + Cz + D = 0$ is

$$d = \frac{|Ax_1 + By_1 + Cz_1 + D|}{\sqrt{A^2 + B^2 + C^2}}$$

Space Curves Space curves are usually specified as the set of points whose coordinates are given parametrically by a system of equations $x = f(t)$, $y = g(t)$, $z = h(t)$ in the parameter t.

Example The equation of a straight line in space is $(x - x_1)/a = (y - y_1)/b = (z - z_1)/c$. Since all these quantities must be equal (say, to t), we may write $x = x_1 + at$, $y = y_1 + bt$, $z = z_1 + ct$, which represent the parametric equations of the line.

Example The equations $z = a\cos\beta t$, $y = a\sin\beta t$, $z = bt$, a, β, b positive constants, represent a circular helix.

Surfaces The locus of points (x, y, z) satisfying $f(x, y, z) = 0$, broadly speaking, may be interpreted as a surface. The simplest surface is the **plane.** The next simplest is a **cylinder,** which is a surface generated by a straight line moving parallel to a given line and passing through a given curve.

Example The parabolic cylinder $y = x^2$ (Fig. 3-28) is generated by a straight line parallel to the z axis passing through $y = x^2$ in the plane $z = 0$.

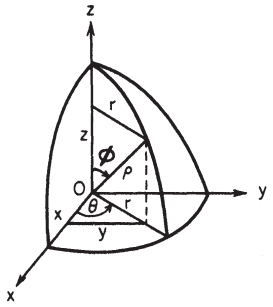

FIG. 3-27 Spherical coordinates.

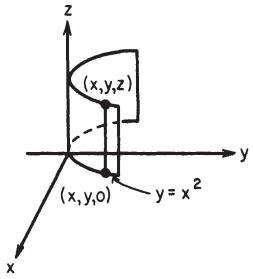

FIG. 3-28 Parabolic cylinder.

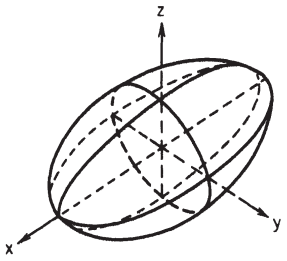

FIG. 3-29 Ellipsoid. $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ (sphere if $a = b = c$)

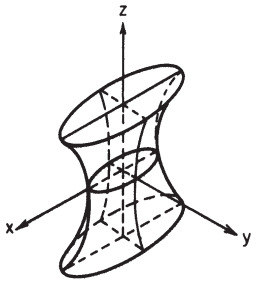

FIG. 3-30 Hyperboloid of one sheet. $\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$

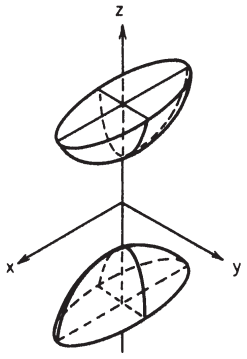

FIG. 3-31 Hyperboloid of two sheets. $\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1$

A surface whose equation is a quadratic in the variables x, y, and z is called a **quadric surface.** Some of the more common such surfaces are tabulated and pictured in Figs. 3-29 to 3-37.

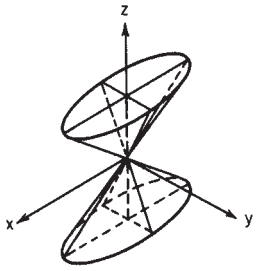

FIG. 3-32 Cone. $\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=0$

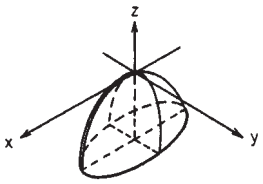

FIG. 3-33 Elliptic paraboloid. $\frac{x^2}{a^2}+\frac{y^2}{b^2}+2z=0$

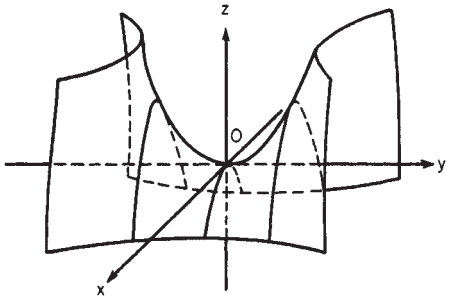

FIG. 3-34 Hyperbolic paraboloid. $\frac{x^2}{a^2}-\frac{y^2}{b^2}+2z=0$

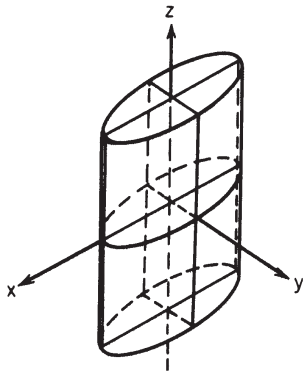

FIG. 3-35 Elliptic cylinder. $\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$

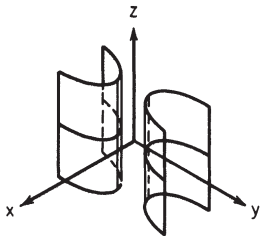

FIG. 3-36 Hyperbolic cylinder. $\frac{x^2}{a^2}-\frac{y^2}{b^2}=1$

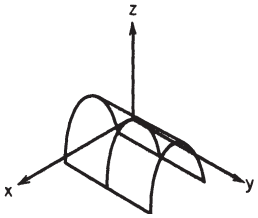

FIG. 3-37 Parabolic cylinder. $y^2+2ax=0$

PLANE TRIGONOMETRY

REFERENCES: 20, 108, 131, 158, 166, 202.

ANGLES

An angle is generated by the rotation of a line about a fixed center from some initial position to some terminal position. If the rotation is clockwise, the angle is negative; if it is counterclockwise, the angle is positive. Angle size is unlimited. If α, β are two angles such that $\alpha+\beta=90°$, they are complementary; they are supplementary if $\alpha+\beta=180°$. Angles are most commonly measured in the sexagesimal system or by radian measure. In the first system there are 360 degrees in one complete revolution; one degree = 1/90 of a right angle. The degree is subdivided into 60 minutes; the minute is subdivided into 60 seconds. In the radian system one radian is the angle at the center of a circle subtended by an arc whose length is equal to the radius of the circle. Thus 2π rad = 360°; 1 rad = 57.29578°; 1° = 0.01745 rad; 1 min = 0.00029089 rad. The advantage of radian measure is that it is *dimensionless.* The quadrants are conventionally labeled as Fig. 3-38 shows.

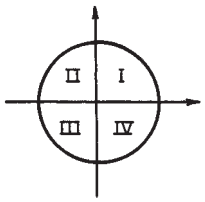

FIG. 3-38 Quadrants.

FUNCTIONS OF CIRCULAR TRIGONOMETRY

The trigonometric functions of angles are the ratios between the various sides of the reference triangles shown in Fig. 3-39 for the various quadrants. Clearly $r=\sqrt{x^2+y^2}\geq 0$. The fundamental functions (see Figs. 3-40, 3-41, 3-42) are

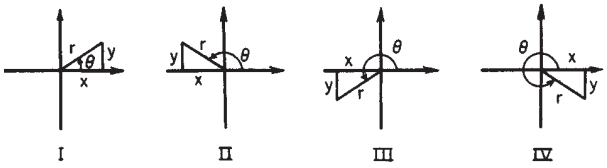

FIG. 3-39 Triangles.

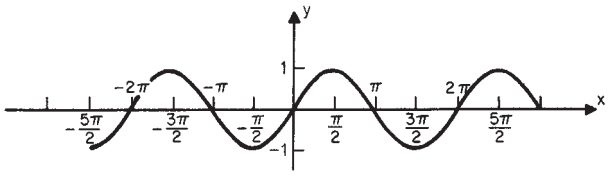

FIG. 3-40 Graph of $y=\sin x$.

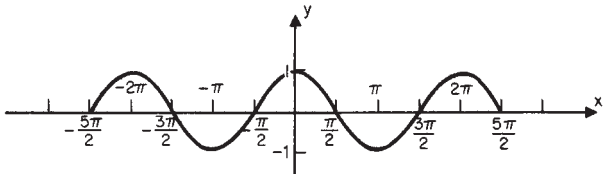

FIG. 3-41 Graph of $y = \cos x$.

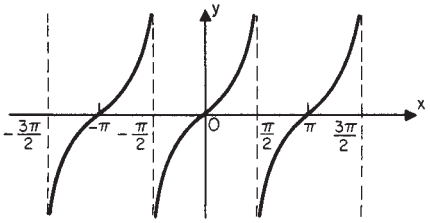

FIG. 3-42 Graph of $y = \tan x$.

Plane Trigonometry

Sine of $\theta = \sin\theta = y/r$	Secant of $\theta = \sec\theta = r/x$
Cosine of $\theta = \cos\theta = x/r$	Cosecant of $\theta = \csc\theta = r/y$
Tangent of $\theta = \tan\theta = y/x$	Cotangent of $\theta = \cot\theta = x/y$

Magnitude and Sign of Trigonometric Functions $0 \le \theta \le 360°$

Function	0° to 90°	90° to 180°	180° to 270°	270° to 360°
$\sin\theta$	+0 to +1	+1 to +0	−0 to −1	−1 to −0
$\csc\theta$	+∞ to +1	+1 to +∞	−∞ to −1	−1 to −∞
$\cos\theta$	+1 to 0	−0 to −1	−1 to −0	+0 to +1
$\sec\theta$	+1 to +∞	−∞ to −1	−1 to −∞	+∞ to +1
$\tan\theta$	+0 to +∞	−∞ to −0	+0 to +∞	−∞ to −0
$\cot\theta$	+∞ to +0	−0 to −∞	+∞ to +0	−0 to −∞

Values of the Trigonometric Functions for Common Angles

θ°	θ, rad	sin θ	cos θ	tan θ
0	0	0	1	0
30	π/6	1/2	$\sqrt{3}/2$	$\sqrt{3}/3$
45	π/4	$\sqrt{2}/2$	$\sqrt{2}/2$	1
60	π/3	$\sqrt{3}/2$	1/2	$\sqrt{3}$
90	π/2	1	0	+∞

If $90° \le \theta \le 180°$, $\sin\theta = \sin(180° - \theta)$; $\cos\theta = -\cos(180° - \theta)$; $\tan\theta = -\tan(180° - \theta)$. If $180° \le \theta \le 270°$, $\sin\theta = -\sin(270° - \theta)$; $\cos\theta = -\cos(270° - \theta)$; $\tan\theta = \tan(270° - \theta)$. If $270° \le \theta \le 360°$, $\sin\theta = -\sin(360° - \theta)$; $\cos\theta = \cos(360° - \theta)$; $\tan\theta = -\tan(360° - \theta)$. The reciprocal properties may be used to find the values of the other functions.

If it is desired to find the angle when a function of it is given, the procedure is as follows: There will in general be two angles between 0° and 360° corresponding to the given value of the function.

Given ($a > 0$)	Find an acute angle θ_0 such that	Required angles are
$\sin\theta = +a$	$\sin\theta_0 = a$	θ_0 and $(180° - \theta_0)$
$\cos\theta = +a$	$\cos\theta_0 = a$	θ_0 and $(360° - \theta_0)$
$\tan\theta = +a$	$\tan\theta_0 = a$	θ_0 and $(180° + \theta_0)$
$\sin\theta = -a$	$\sin\theta_0 = a$	$180° + \theta_0$ and $360° - \theta_0$
$\cos\theta = -a$	$\cos\theta_0 = a$	$180° - \theta_0$ and $180° + \theta_0$
$\tan\theta = -a$	$\tan\theta_0 = a$	$180° - \theta_0$ and $360° - \theta_0$

Relations between Functions of a Single Angle $\sec\theta = 1/\cos\theta$; $\csc\theta = 1/\sin\theta$, $\tan\theta = \sin\theta/\cos\theta = \sec\theta/\csc\theta = 1/\cot\theta$; $\sin^2\theta + \cos^2\theta = 1$; $1 + \tan^2\theta = \sec^2\theta$; $1 + \cot^2\theta = \csc^2\theta$. For $0 \le \theta \le 90°$ the following results hold:

$$\sin\theta = \cos\theta/\cot\theta = \sqrt{1 - \cos^2\theta} = \cos\theta\tan\theta$$

$$= \frac{\tan\theta}{\sqrt{1 + \tan^2\theta}} = \frac{1}{\sqrt{1 + \cot^2\theta}} = 2\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)$$

and $$\cos\theta = \sqrt{1 - \sin^2\theta} = \frac{1}{\sqrt{1 + \tan^2\theta}}$$

$$= \frac{\cot\theta}{\sqrt{1 + \cot^2\theta}} = \frac{\sin\theta}{\tan\theta} = \cos^2\left(\frac{\theta}{2}\right) - \sin^2\left(\frac{\theta}{2}\right)$$

The cofunction property is very important. $\cos\theta = \sin(90° - \theta)$, $\sin\theta = \cos(90° - \theta)$, $\tan\theta = \cot(90° - \theta)$, $\cot\theta = \tan(90° - \theta)$, etc.

Functions of Negative Angles $\sin(-\theta) = -\sin\theta$, $\cos(-\theta) = \cos\theta$, $\tan(-\theta) = -\tan\theta$, $\sec(-\theta) = \sec\theta$, $\csc(-\theta) = -\csc\theta$, $\cot(-\theta) = -\cot\theta$.

Identities

Sum and Difference Formulas Let x, y be two angles. $\sin(x \pm y) = \sin x\cos y \pm \cos x\sin y$; $\cos(x \pm y) = \cos x\cos y \mp \sin x\sin y$; $\tan(x \pm y) = (\tan x \pm \tan y)/(1 \mp \tan x\tan y)$; $\sin x \pm \sin y = 2\sin\frac{1}{2}(x \pm y)\cos\frac{1}{2}(x \mp y)$; $\cos x + \cos y = 2\cos\frac{1}{2}(x + y)\cos\frac{1}{2}(x - y)$; $\cos x - \cos y = -2\sin\frac{1}{2}(x + y)\sin\frac{1}{2}(x - y)$; $\tan x \pm \tan y = [\sin(x \pm y)]/(\cos x\cos y)$; $\sin^2 x - \sin^2 y = \cos^2 y - \cos^2 x = \sin(x + y)\sin(x - y)$; $\cos^2 x - \sin^2 y = \cos^2 y - \sin^2 x = \cos(x + y)\cos(x - y)$; $\sin(45° + x) = \cos(45° - x)$; $\sin(45° - x) = \cos(45° + x)$; $\tan(45° \pm x) = \cot(45° \mp x)$. $A\cos x + B\sin x = \sqrt{A^2 + B^2}\sin(\alpha + x) = \sqrt{A^2 + B^2}\cos(\beta - x)$ where $\tan\alpha = A/B$, $\tan\beta = B/A$; both α and β are positive acute angles.

Multiple and Half Angle Identities Let x = angle, $\sin 2x = 2\sin x\cos x$; $\sin x = 2\sin\frac{1}{2}x\cos\frac{1}{2}x$; $\cos 2x = \cos^2 x - \sin^2 x = 1 - 2\sin^2 x = 2\cos^2 x - 1$. $\tan 2x = (2\tan x)/(1 - \tan^2 x)$; $\sin 3x = 3\sin x - 4\sin^3 x$; $\cos 3x = 4\cos^3 x - 3\cos x$. $\tan 3x = (3\tan x - \tan^3 x)/(1 - 3\tan^2 x)$; $\sin 4x = 4\sin x\cos x - 8\sin^3 x\cos x$; $\cos 4x = 8\cos^4 x - 8\cos^2 x + 1$.

$$\sin\left(\frac{x}{2}\right) = \sqrt{\tfrac{1}{2}(1 - \cos x)}$$

$$\cos\left(\frac{x}{2}\right) = \sqrt{\tfrac{1}{2}(1 + \cos x)}$$

$$\tan\left(\frac{x}{2}\right) = \sqrt{\frac{1 - \cos x}{1 + \cos x}} = \frac{\sin x}{1 + \cos x} = \frac{1 - \cos x}{\sin x}$$

Relations between Three Angles Whose Sum Is 180° Let x, y, z be the angles.

$$\sin x + \sin y + \sin z = 4\cos\left(\frac{x}{2}\right)\cos\left(\frac{y}{2}\right)\cos\left(\frac{z}{2}\right)$$

$$\cos x + \cos y + \cos z = 4\sin\left(\frac{x}{2}\right)\sin\left(\frac{y}{2}\right)\sin\left(\frac{z}{2}\right) + 1$$

$$\sin x + \sin y - \sin z = 4\sin\left(\frac{x}{2}\right)\sin\left(\frac{y}{2}\right)\cos\left(\frac{z}{2}\right)$$

$\sin^2 x + \sin^2 y + \sin^2 z = 2\cos x\cos y\cos z + 2$; $\tan x + \tan y + \tan z = \tan x\tan y\tan z$; $\sin 2x + \sin 2y + \sin 2z = 4\sin x\sin y\sin z$.

INVERSE TRIGONOMETRIC FUNCTIONS

$y = \sin^{-1} x = \arcsin x$ is the angle y whose sine is x.

Example $y = \sin^{-1}\frac{1}{2}$, y is 30°.

The complete solution of the equation $x = \sin y$ is $y = (-1)^n\sin^{-1} x + n(180°)$, $-\pi/2 \le \sin^{-1} x \le \pi/2$ where $\sin^{-1} x$ is the principal value of the angle whose sine is x. The range of principal values of the $\cos^{-1} x$ is $0 \le \cos^{-1} x \le \pi$ and $-\pi/2 \le \tan^{-1} x \le \pi/2$. If these restrictions are allowed to hold, the following formulas result:

$$\sin^{-1} x = \cos^{-1}\sqrt{1-x^2} = \tan^{-1}\frac{x}{\sqrt{1-x^2}} = \cot^{-1}\frac{\sqrt{1-x^2}}{x}$$

$$= \sec^{-1}\frac{1}{\sqrt{1-x^2}} = \csc^{-1}\frac{1}{x} = \frac{\pi}{2} - \cos^{-1} x$$

$$\cos^{-1} x = \sin^{-1}\sqrt{1-x^2} = \tan^{-1}\frac{\sqrt{1-x^2}}{x}$$

$$= \cot^{-1}\frac{x}{\sqrt{1-x^2}} = \sec^{-1}\frac{1}{x}$$

$$= \csc^{-1}\frac{1}{\sqrt{1-x^2}} = \frac{\pi}{2} - \sin^{-1} x$$

$$\tan^{-1} x = \sin^{-1}\frac{x}{\sqrt{1-x^2}} = \cos^{-1}\frac{1}{\sqrt{1+x^2}}$$

$$= \cot^{-1}\frac{1}{x} = \sec^{-1}\sqrt{1+x^2} = \csc^{-1}\frac{\sqrt{1+x^2}}{x}$$

RELATIONS BETWEEN ANGLES AND SIDES OF TRIANGLES

Solutions of Triangles (Fig. 3-43) Let a, b, c denote the sides and α, β, γ the angles opposite the sides in the triangle. Let $2s = a + b + c$, A = area, r = radius of the inscribed circle, R = radius of the circumscribed circle, and h = altitude. In any triangle $\alpha + \beta + \gamma = 180°$.

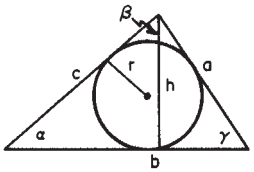

FIG. 3-43 Triangle.

Law of Sines $\sin\alpha/a = \sin\beta/b = \sin\gamma/c$.

Law of Tangents

$$\frac{a+b}{a-b} = \frac{\tan ½(\alpha+\beta)}{\tan ½(\alpha-\beta)};\quad \frac{b+c}{b-c} = \frac{\tan ½(\beta+\gamma)}{\tan ½(\beta-\gamma)};\quad \frac{a+c}{a-c} = \frac{\tan ½(\alpha+\gamma)}{\tan ½(\alpha-\gamma)}$$

Law of Cosines $a^2 = b^2 + c^2 - 2bc\cos\alpha$; $b^2 = a^2 + c^2 - 2ac\cos\beta$; $c^2 = a^2 + b^2 - 2ab\cos\gamma$.

Other Relations In this subsection, where appropriate, two more formulas can be generated by replacing a by b, b by c, c by a, α by β, β by γ, and γ by α. $\cos\alpha = (b^2 + c^2 - a^2)/2bc$; $a = b\cos\gamma + c\cos\beta$; $\sin\alpha = (2/bc)\sqrt{s(s-a)(s-b)(s-c)}$;

$$\sin\left(\frac{\alpha}{2}\right) = \sqrt{\frac{(s-b)(s-c)}{bc}};\quad \cos\left(\frac{\alpha}{2}\right) = \sqrt{\frac{s(s-a)}{bc}};\quad A = \frac{1}{2}bh$$

$$= \frac{1}{2}ab\sin\gamma = \frac{a^2\sin\beta\sin\gamma}{2\sin\alpha} = \sqrt{s(s-a)(s-b)(s-c)} = rs$$

where $r = \sqrt{\dfrac{(s-a)(s-b)(s-c)}{s}}$

$$R = a/(2\sin\alpha) = abc/4A;\quad h = c\sin\alpha = a\sin\gamma = 2rs/b.$$

Example $a = 5$, $b = 4$, $\alpha = 30°$. Use the law of sines. $0.5/5 = \sin\beta/4$, $\sin\beta = ⅖$, $\beta = 23°35'$, $\gamma = 126°25'$. So $c = \sin 126°25'/⅒ = 10(.8047) = 8.05$.

The relations given here suffice to solve any triangle. One method for each triangle is given.

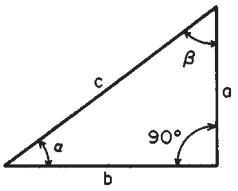

FIG. 3-44 Right triangle.

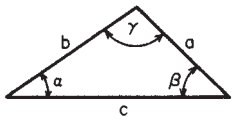

FIG. 3-45 Oblique triangle.

Right Triangle (Fig. 3-44) Given one side and any acute angle α or any two sides, the remaining parts can be obtained from the following formulas:

$$a = \sqrt{(c+b)(c-b)} = c\sin\alpha = b\tan\alpha$$

$$b = \sqrt{(c+a)(c-a)} = c\cos\alpha = a\cot\alpha$$

$$c = \sqrt{a^2 \pm b^2},\ \sin\alpha = \frac{a}{c},\ \cos\alpha = \frac{b}{c},\ \tan\alpha = \frac{a}{b},\ \beta = 90° - \alpha$$

$$A = \frac{1}{2}ab = \frac{a^2}{2\tan\alpha} = \frac{b^2\tan\alpha}{2} = \frac{c^2\sin 2\alpha}{4}$$

Oblique Triangles (Fig. 3-45) There are four possible cases.

1. Given b, c and the included angles α,

$$\frac{1}{2}(\beta+\gamma) = 90° - \frac{1}{2}\alpha;\ \tan\frac{1}{2}(\beta-\gamma) = \frac{b-c}{b+c}\tan\frac{1}{2}(\beta+\gamma)$$

$$\beta = \frac{1}{2}(\beta+\gamma) + \frac{1}{2}(\beta-\gamma);\ \gamma = \frac{1}{2}(\beta+\gamma) - \frac{1}{2}(\beta-\gamma);\ a = \frac{b\sin\alpha}{\sin\beta}$$

2. Given the three sides a, b, c, $s = ½(a+b+c)$;

$$r = \sqrt{\frac{(s-a)(s-b)(s-c)}{s}}$$

$$\tan\frac{1}{2}\alpha = \frac{r}{s-a};\ \tan\frac{1}{2}\beta = \frac{r}{s-b};\ \tan\frac{1}{2}\gamma = \frac{r}{s-c}$$

3. Given any two sides a, c and an angle opposite one of them α, $\sin\gamma = (c\sin\alpha)/a$; $\beta = 180° - a - \gamma$; $b = (a\sin\beta)/(\sin\alpha)$. There may be two solutions here. γ may have two values γ_1, γ_2; $\gamma_1 < 90°$, $\gamma_2 = 180° - \gamma_1 > 90°$. If $\alpha + \gamma_2 > 180°$, use only γ_1. This case may be *impossible* if $\sin\gamma > 1$.

4. Given any side c and two angles α and β, $\gamma = 180° - \alpha - \beta$; $a = (c\sin\alpha)/(\sin\gamma)$; $b = (c\sin\beta)/(\sin\gamma)$.

HYPERBOLIC TRIGONOMETRY

The hyperbolic functions are certain combinations of exponentials e^x and e^{-x}.

$$\cosh x = \frac{e^x + e^{-x}}{2};\ \sinh x = \frac{e^x - e^{-x}}{2};\ \tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\coth x = \frac{e^x + e^{-x}}{e^x - e^{-x}} = \frac{1}{\tanh x} = \frac{\cosh x}{\sinh x};\ \operatorname{sech} x = \frac{1}{\cosh x} = \frac{2}{e^x + e^{-x}};$$

$$\operatorname{csch} x = \frac{1}{\sinh x} = \frac{2}{e^x - e^{-x}}$$

Fundamental Relationships $\sinh x + \cosh x = e^x$; $\cosh x - \sinh x = e^{-x}$; $\cosh^2 x - \sinh^2 x = 1$; $\operatorname{sech}^2 x + \tanh^2 x = 1$; $\coth^2 x - \operatorname{csch}^2 x = 1$; $\sinh 2x = 2\sinh x\cosh x$; $\cosh 2x = \cosh^2 x + \sinh^2 x = 1 + 2\sinh^2 x = 2\cosh^2 x - 1$. $\tanh 2x = (2\tanh x)/(1 + \tanh^2 x)$; $\sinh(x \pm y) = \sinh x\cosh y \pm \cosh x\sinh y$; $\cosh(x \pm y) = \cosh x\cosh y \pm \sinh x\sinh y$; $2\sinh^2 x/2 = \cosh x - 1$; $2\cosh^2 x/2 = \cosh x + 1$; $\sinh(-x) = -\sinh x$; $\cosh(-x) = \cosh x$; $\tanh(-x) = -\tanh x$.

When $u = a\cosh x$, $v = a\sinh x$, then $u^2 - v^2 = a^2$; which is the equation for a hyperbola. In other words, the hyperbolic functions in the parametric equations $u = a\cosh x$, $v = a\sinh x$ have the same relation to the hyperbola $u^2 - v^2 = a^2$ that the equations $u = a\cos\theta$, $v = a\sin\theta$ have to the circle $u^2 + v^2 = a^2$.

Inverse Hyperbolic Functions If $x = \sinh y$, then y is the inverse hyperbolic sine of x written $y = \sinh^{-1} x$ or $\operatorname{arcsinh} x$. $\sinh^{-1} x = \log_e (x + \sqrt{x^2 + 1})$

$$\cosh^{-1} x = \log_e (x + \sqrt{x^2 + 1}); \quad \tanh^{-1} x = \frac{1}{2} \log_e \frac{1 + x}{1 - x};$$

$$\coth^{-1} x = \frac{1}{2} \log_e \frac{x + 1}{x - 1}; \quad \operatorname{sech}^{-1} x = \log_e \left(\frac{1 + \sqrt{1 - x^2}}{x} \right);$$

$$\operatorname{csch}^{-1} = \log_e \left(\frac{1 + \sqrt{1 + x^2}}{x} \right)$$

Magnitude of the Hyperbolic Functions $\cosh x \geq 1$ with equality only for $x = 0$; $-\infty < \sinh x < \infty$; $-1 < \tanh x < 1$. $\cosh x \sim e^x/2$ as $x \to \infty$; $\sinh x \to e^x/2$ as $x \to \infty$.

APPROXIMATIONS FOR TRIGONOMETRIC FUNCTIONS

For small values of θ (θ measured in radians) $\sin \theta \approx \theta$, $\tan \theta \approx \theta$; $\cos \theta \approx 1 - (\theta^2/2)$. The following relations actually hold: $\sin \theta < \theta < \tan \theta$; $\cos \theta < \sin \theta/\theta < 1$; $\theta \sqrt{1 - \theta^2} < \sin \theta < \theta$; $\cos \theta < \theta/\tan \theta < 1$;

$$\theta \left(1 - \frac{\theta^2}{2} \right) < \sin \theta < \theta \text{ and } \theta < \tan \theta < \frac{\theta}{\sqrt{1 - \theta^2}}$$

The behavior ratio of the functions as $\theta \to 0$ is given by the following:

$$\lim_{\theta \to 0} \sin \theta/\theta = 1; \quad \sin \theta/\tan \theta = 1.$$

DIFFERENTIAL AND INTEGRAL CALCULUS

REFERENCES: 114, 158, 260, 261, 274, 282, 296. See also "General References: References for General and Specific Topics—Advanced Calculus." For computer evaluations of the calculus described here, see Refs. 68, 299.

DIFFERENTIAL CALCULUS

An Example of Functional Notation Suppose that a storage warehouse of 16,000 ft^3 is required. The construction costs per square foot are \$10, \$3, and \$2 for walls, roof, and floor respectively. What are the minimum cost dimensions? Thus, with h = height, x = width, and y = length, the respective costs are

$$\text{Walls} = 2 \times 10hy + 2 \times 10hx = 20h(y + x)$$
$$\text{Roof} = 3xy$$
$$\text{Floor} = 2xy$$
$$\text{Total cost} = 2xy + 3xy + 20h(x + y) = 5xy + 20h(x + y) \qquad (3\text{-}1)$$

and the restriction

$$\text{Total volume} = xyh \qquad (3\text{-}2)$$

Solving for h from Eq. (3-2),

$$h = \text{volume}/xy = 16{,}000/xy \qquad (3\text{-}3)$$

$$\text{Cost} = 5xy + \frac{320{,}000}{xy} (y + x) = 5xy + 320{,}000 \left(\frac{1}{x} + \frac{1}{y} \right) \qquad (3\text{-}4)$$

In this form it can be shown that the minimum cost will occur for $x = y$; therefore

$$\text{Cost} = 5x^2 + 640{,}000\ (1/x)$$

By evaluation, the smallest cost will occur when $x = 40$.

$$\text{Cost} = 5(1600) + 640{,}000/40 = \$24{,}000$$

The dimensions are then $x = 40$ ft, $y = 40$ ft, $h = 16{,}000/(40 \times 40) = 10$ ft. Symbolically, the original cost relationship is written

$$\text{Cost} = f(x, y, h) = 5xy + 20h(y + x)$$

and the volume relation

$$\text{Volume} = g(x, y, h) = xyh = 16{,}000$$

In terms of the derived general relationships (3-1) and (3-2), x, y, and h are **independent variables—**cost and volume, **dependent variables.** That is, the cost and volume become fixed with the specification of dimensions. However, corresponding to the given restriction of the problem, relative to volume, the function $g(x, y, z) = xyh$ becomes a constraint function. In place of three independent and two dependent variables the problem reduces to two independent (volume has been constrained) and two dependent as in functions (3-3) and (3-4). Further, the requirement of minimum cost reduces the problem to three dependent variables (x, y, h) and no **degrees of freedom,** that is, freedom of independent selection.

Limits The limit of function $f(x)$ as x approaches a (a is finite or else x is said to increase without bound) is the number N.

$$\lim_{x \to a} f(x) = N$$

This states that $f(x)$ can be calculated as close to N as desirable by making x sufficiently close to a. This does not put any restriction on $f(x)$ when $x = a$. Alternatively, for any given positive number ε, a number δ can be found such that $0 < |a - x| < \delta$ implies that $|N - f(x)| < \varepsilon$.

The following operations with limits (when they exist) are valid:

$$\lim_{x \to a} bf(x) = b \lim_{x \to a} f(x)$$

$$\lim_{x \to a} [f(x) + g(x)] = \lim_{x \to a} f(x) + \lim_{x \to a} g(x)$$

$$\lim_{x \to a} [f(x)g(x)] = \lim_{x \to a} f(x) \cdot \lim_{x \to a} g(x)$$

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{\lim_{x \to a} f(x)}{\lim_{x \to a} g(x)} \quad \text{if} \quad \lim_{x \to a} g(x) \neq 0$$

Continuity A function $f(x)$ is continuous at the point $x = a$ if

$$\lim_{h \to 0} [f(a + h) - f(a)] = 0$$

Rigorously, it is stated $f(x)$ is continuous at $x = a$ if for any positive ε there exists a $\delta > 0$ such that $|f(a + h) - f(a)| < \varepsilon$ for all x with $|x - a| < \delta$. For example, the function $(\sin x)/x$ is not continuous at $x = 0$ and therefore is said to be discontinuous. Discontinuities are classified into three types:

1.	Removable	$y = \sin x/x$	at $x = 0$
2.	Infinite	$y = 1/x$	at $x = 0$
3.	Jump	$y = 10/(1 + e^{1/x})$	at $x = 0^+$ $y = 0^+$
			$x = 0$ $y = 0$
			$x = 0^-$ $y = 10$

Derivative The function $f(x)$ has a derivative at $x = a$, which can be denoted as $f'(a)$, if

$$\lim_{h \to 0} \frac{f(a + h) - f(a)}{h}$$

exists. This implies continuity at $x = a$. Conversely, a function may be continuous but not have a derivative. The derivative function is

$$f'(x) = \frac{df}{dx} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

Differentiation Define $\Delta y = f(x + \Delta x) - f(x)$. Then dividing by Δx

$$\frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Call $$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \frac{dy}{dx}$$

then $$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Example Find the derivative of $y = \sin x$.

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\sin(x + \Delta x) - \sin(x)}{\Delta x}$$

$$= \lim_{\Delta x \to 0} \frac{\sin x \cos \Delta x + \sin \Delta x \cos x - \sin x}{\Delta x}$$

$$= \lim_{\Delta x \to 0} \frac{\sin x(\cos \Delta x - 1)}{\Delta x} + \lim_{\Delta x \to 0} \frac{\sin \Delta x \cos x}{\Delta x}$$

$$= \cos x \text{ since } \lim_{\Delta x \to 0} \frac{\sin \Delta x}{\Delta x} = 1$$

Differential Operations The following differential operations are valid: f, g, . . . are differentiable functions of x, c is a constant; e is the base of the natural logarithms.

$$\frac{dc}{dx} = 0 \tag{3-5}$$

$$\frac{dx}{dx} = 1 \tag{3-6}$$

$$\frac{d}{dx}(f + g) = \frac{df}{dx} + \frac{dg}{dx} \tag{3-7}$$

$$\frac{d}{dx}(f \times g) = f\frac{dg}{dx} + g\frac{df}{dx} \tag{3-8}$$

$$\frac{dy}{dx} = \frac{1}{dx/dy} \quad \text{if} \quad \frac{dx}{dy} \neq 0 \tag{3-9}$$

$$\frac{d}{dx} f^n = nf^{n-1}\frac{df}{dx} \tag{3-10}$$

$$\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{g(df/dx) - f(dg/dx)}{g^2} \tag{3-11}$$

$$\frac{df}{dx} = \frac{df}{dv} \times \frac{dv}{dx} \quad \text{(chain rule)} \tag{3-12}$$

$$\frac{df^g}{dx} = gf^{g-1}\frac{df}{dx} + f^g \ln f \frac{dg}{dx} \tag{3-13}$$

$$\frac{da^x}{dx} = (\ln a)\, a^x \tag{3-14}$$

Example Derive dy/dx for $x^2 + y^3 = x + xy + A$.

Here $$\frac{d}{dx}x^2 + \frac{d}{dx}y^3 = \frac{d}{dx}x + \frac{d}{dx}xy + \frac{d}{dx}A$$

$$2x + 3y^2\frac{dy}{dx} = 1 + y + x\frac{dy}{dx} + 0$$

by rules (3-10), (3-10), (3-6), (3-8), and (3-5) respectively.

Thus $$\frac{dy}{dx} = \frac{2x - 1 - y}{x - 3y^2}$$

Differentials

$$de^x = e^x\, dx \tag{3-15a}$$

$$d(a^x) = a^x \log a\, dx \tag{3-15b}$$

$$d \ln x = (1/x)\, dx \tag{3-16}$$

$$d \log x = (\log e/x)dx \tag{3-17}$$

$$d \sin x = \cos x\, dx \tag{3-18}$$

$$d \cos x = -\sin x\, dx \tag{3-19}$$

$$d \tan x = \sec^2 x\, dx \tag{3-20}$$

$$d \cot x = -\csc^2 x\, dx \tag{3-21}$$

$$d \sec x = \tan x \sec x\, dx \tag{3-22}$$

$$d \csc x = -\cot x \csc x\, dx \tag{3-23}$$

$$d \sin^{-1} x = (1 - x^2)^{-1/2}\, dx \tag{3-24}$$

$$d \cos^{-1} x = -(1 - x^2)^{-1/2}\, dx \tag{3-25}$$

$$d \tan^{-1} x = (1 + x^2)^{-1}\, dx \tag{3-26}$$

$$d \cot^{-1} x = -(1 + x^2)^{-1}\, dx \tag{3-27}$$

$$d \sec^{-1} x = x^{-1}(x^2 - 1)^{-1/2}\, dx \tag{3-28}$$

$$d \csc^{-1} x = -x^{-1}(x^2 - 1)^{-1/2}\, dx \tag{3-29}$$

$$d \sinh x = \cosh x\, dx \tag{3-30}$$

$$d \cosh x = \sinh x\, dx \tag{3-31}$$

$$d \tanh x = \operatorname{sech}^2 x\, dx \tag{3-32}$$

$$d \coth x = -\operatorname{csch}^2 x\, dx \tag{3-33}$$

$$d \operatorname{sech} x = -\operatorname{sech} x \tanh x\, dx \tag{3-34}$$

$$d \operatorname{csch} x = -\operatorname{csch} x \coth x\, dx \tag{3-35}$$

$$d \sinh^{-1} x = (x^2 + 1)^{-1/2}\, dx \tag{3-36}$$

$$d \cosh^{-1} = (x^2 - 1)^{-1/2}\, dx \tag{3-37}$$

$$d \tanh^{-1} x = (1 - x^2)^{-1}\, dx \tag{3-38}$$

$$d \coth^{-1} x = -(x^2 - 1)^{-1}\, dx \tag{3-39}$$

$$d \operatorname{sech}^{-1} x = -(1/x)(1 - x^2)^{-1/2}\, dx \tag{3-40}$$

$$d \operatorname{csch}^{-1} x = -x^{-1}(x^2 + 1)^{-1/2}\, dx \tag{3-41}$$

Example Find dy/dx for $y = \sqrt{x} \cos(1 - x^2)$.

Using

$$\frac{dy}{dx} = \sqrt{x}\frac{d}{dx}\cos(1 - x^2) + \cos(1 - x^2)\frac{d}{dx}\sqrt{x} \tag{3-8}$$

$$\frac{d}{dx}\cos(1 - x^2) = -\sin(1 - x^2)\frac{d}{dx}(1 - x^2) \tag{3-19}$$

$$= -\sin(1 - x^2)(0 - 2x) \tag{3-5), (3-10}$$

$$\frac{d\sqrt{x}}{dx} = \frac{1}{2}x^{-1/2} \tag{3-10}$$

$$\frac{dy}{dx} = 2x^{3/2}\sin(1 - x^2) + \frac{1}{2}x^{-1/2}\cos(1 - x^2)$$

Example Find the derivative of $\tan x$ with respect to $\sin x$.

$$v = \sin x$$

$$y = \tan x$$

Using

$$\frac{d \tan x}{d \sin x} = \frac{dy}{dv} = \frac{dy}{dx}\frac{dx}{dv} \tag{3-12}$$

$$= \frac{d \tan x}{dx}\frac{1}{\dfrac{d \sin x}{dx}} \tag{3-9}$$

$$= \sec^2 x/\cos x \tag{3-18), (3-20}$$

Very often in experimental sciences and engineering functions and their derivatives are available only through their numerical values. In particular, through measurements we may know the values of a function and its derivative only at certain points. In such cases the preceding operational rules for derivatives, including the chain rule, can be applied numerically.

Example Given the following table of values for differentiable functions f and g; evaluate the following quantities:

x	$f(x)$	$f'(x)$	$g(x)$	$g'(x)$
1	3	1	4	−4
3	0	2	4	7
4	−2	10	3	6

$$\frac{d}{dx}[f(x)+g(x)]|_{x=4} = f'(4)+g'(4) = 10+6 = 16$$

$$\left(\frac{f}{g}\right)'(1) = \frac{f'(1)g(1)-f(1)g'(1)}{[g(1)]^2} = \frac{1\cdot 4-3(-4)}{(-4)^2} = \frac{16}{16} = 1$$

Higher Differentials The first derivative of $f(x)$ with respect to x is denoted by f' or df/dx. The derivative of the first derivative is called the second derivative of $f(x)$ with respect to x and is denoted by f'', $f^{(2)}$, or d^2f/dx^2; and similarly for the higher-order derivatives.

Example Given $f(x) = 3x^3 + 2x + 1$, calculate all derivative values at $x = 3$.

$$\frac{df(x)}{dx} = 9x^2+2 \qquad x=3, f'(3) = 9(9)+2 = 83$$

$$\frac{d^2f(x)}{dx^2} = 18x \qquad x=3, f''(3) = 18(3) = 54$$

$$\frac{d^3f(x)}{dx^3} = 18 \qquad x=3, f''(3) = 18$$

$$\frac{d^nf(x)}{dx^n} = 0 \qquad \text{for } n \geq 4$$

If $f'(x) > 0$ on (a, b), then f is increasing on (a, b). If $f'(x) < 0$ on (a, b), then f is decreasing on (a, b).

The graph of a function $y = f(x)$ is concave up if f' is increasing on (a, b); it is concave down if f' is decreasing on (a, b).

If $f''(x)$ exists on (a, b) and if $f''(x) > 0$, then f is concave up on (a, b). If $f''(x) < 0$, then f is concave down on (a, b).

An inflection point is a point at which a function changes the direction of its concavity.

Indeterminate Forms: L'Hospital's Theorem Forms of the type 0/0, ∞/∞, $0 \times \infty$, etc., are called indeterminates. To find the limiting values that the corresponding functions approach, L'Hospital's theorem is useful: If two functions $f(x)$ and $g(x)$ both become zero at $x = a$, then the limit of their quotient is equal to the limit of the quotient of their separate derivatives, if the limit exists or is $+\infty$ or $-\infty$.

Example Find $\lim_{n\to 0} \frac{\sin x}{x}$.

Here $$\lim_{x\to 0}\frac{\sin x}{x} = \lim_{x\to 0}\frac{d\sin x}{dx} = \lim_{x\to 0}\frac{\cos x}{1} = 1$$

Example Find $\lim_{x\to\infty}\frac{(1.1)^x}{x^{1000}}$.

$$\lim_{x\to\infty}\frac{(1.1)^x}{x^{1000}} = \lim_{x\to\infty}\frac{d(1.1)^x}{dx^{1000}} = \lim_{x\to\infty}\frac{(\ln 1.1)(1.1)^x}{1000x^{999}}$$

Obviously $\lim_{x\to\infty}\frac{1.1^x}{x^{1000}} = \infty$ since repeated application of the rule will reduce the denominator to a finite number 1000! while the numerator remains infinitely large.

Example Find $\lim_{x\to\infty} x^3 e^{-x}$.

$$\lim_{x\to\infty} x^3 e^{-x} = \lim_{x\to\infty}\frac{x^3}{e^x} = \lim_{x\to\infty}\frac{6}{e^x} = 0$$

Example Find $\lim_{x\to 0}(1-x)^{1/x}$.

Let $$y = (1-x)^{1/x}$$

$$\ln y = (1/x)\ln(1-x)$$

$$\lim_{x\to 0}(\ln y) = \lim_{x\to 0}\frac{\ln(1-x)}{x} = -1$$

Therefore, $$\lim_{x\to 0} y = e^{-1}$$

Partial Derivative The abbreviation $z = f(x, y)$ means that z is a function of the two variables x and y. The derivative of z with respect to x, treating y as a constant, is called the partial derivative with respect to x and is usually denoted as $\partial z/\partial x$ or $\partial f(x, y)/\partial x$ or simply f_x. Partial differentiation, like full differentiation, is quite simple to apply. Conversely, the solution of partial differential equations is appreciably more difficult than that of differential equations.

Example Find $\partial z/\partial x$ and $\partial z/\partial y$ for $z = ye^{x^2} + xe^y$.

$$\frac{\partial z}{\partial x} = y\frac{\partial e^{x^2}}{\partial x} + e^y\frac{\partial x}{\partial x} \qquad \frac{\partial z}{\partial y} = e^{x^2}\frac{\partial y}{\partial y} + x\frac{\partial e^y}{\partial y}$$

$$= 2xye^{x^2} + e^y \qquad = e^{x^2} + xe^y$$

Order of Differentiation It is generally true that the order of differentiation is immaterial for any number of differentiations or variables provided the function and the appropriate derivatives are continuous. For $z = f(x, y)$ it follows:

$$\frac{\partial^3 f}{\partial y^2\,\partial x} = \frac{\partial^3 f}{\partial y\,\partial x\,\partial y} = \frac{\partial^3 f}{\partial x\,\partial y^2}$$

General Form for Partial Differentiation

1. Given $f(x, y) = 0$ and $x = g(t)$, $y = h(t)$.

Then $$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

$$\frac{d^2f}{dt^2} = \frac{\partial^2 f}{\partial x^2}\left(\frac{dx}{dt}\right)^2 + 2\frac{\partial^2 f}{\partial x\,\partial y}\frac{dx}{dt}\frac{dy}{dt} + \frac{\partial^2 f}{\partial y^2}\left(\frac{dy}{dt}\right)^2 + \frac{\partial f}{\partial x}\frac{d^2x}{dt^2} + \frac{\partial f}{\partial y}\frac{d^2y}{dt^2}$$

Example Find df/dt for $f = xy$, $x = \rho\sin t$, $y = \rho\cos t$.

$$\frac{df}{dt} = \frac{\partial(xy)}{\partial x}\left(\frac{d\,\rho\sin t}{dt}\right) + \frac{\partial(xy)}{\partial y}\left(\frac{d\,\rho\cos t}{dt}\right)$$

$$= y(\rho\cos t) + x(-\rho\sin t)$$

$$= \rho^2\cos^2 t - \rho^2\sin^2 t$$

2. Given $f(x, y) = 0$ and $x = g(t, s)$, $y = h(t, s)$.

Then $$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t}$$

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial x}$$

Differentiation of Composite Function

Rule 1. Given $f(x, y) = 0$, then $\frac{dy}{dx} = -\frac{\partial f/\partial x}{\partial f/\partial y}\left(\frac{\partial f}{\partial y} \neq 0\right)$.

Rule 2. Given $f(u) = 0$ where $u = g(x)$, then

$$\frac{df}{dx} = f'(u)\frac{du}{dx}$$

$$\frac{d^2f}{dx^2} = f''(u)\left(\frac{du}{dx}\right)^2 + f'(u)\frac{d^2u}{dx^2}$$

Example Find df/dx for $f = \sin^2 u$ and $u = \sqrt{1-x^2}$

$$\frac{df}{dx} = \frac{d\sin^2 u}{du}\frac{d\sqrt{1-x^2}}{dx}$$

$$= 2\sin u\cos u\left(\frac{1}{2}\right)(-2x)(1-x^2)^{-1/2}$$

$$= -2\frac{\sqrt{1-u^2}}{u}\sin u\cos u$$

Rule 3. Given $f(u) = 0$ where $u = g(x,y)$, then

$$\frac{\partial f}{\partial x} = f'(u)\frac{\partial u}{\partial x} \qquad \frac{\partial f}{\partial y} = f'(u)\frac{\partial u}{\partial y}$$

$$\frac{\partial^2 f}{\partial x^2} = f''\left(\frac{\partial u}{\partial x}\right)^2 + f'\frac{\partial^2 u}{\partial x^2}$$

$$\frac{\partial^2 f}{\partial x\,\partial y} = f''\frac{\partial u}{\partial x}\frac{\partial u}{\partial y} + f'\frac{\partial^2 u}{\partial x\,\partial y}$$

$$\frac{\partial^2 f}{\partial y^2} = f''\left(\frac{\partial u}{\partial y}\right)^2 + f'\frac{\partial^2 u}{\partial y^2}$$

MULTIVARIABLE CALCULUS APPLIED TO THERMODYNAMICS

Many of the functional relationships needed in thermodynamics are direct applications of the rules of multivariable calculus. This section reviews those rules in the context of the needs of themodynamics. These ideas were expounded in one of the classic books on chemical engineering thermodynamics.[151]

State Functions State functions depend only on the state of the system, not on past history or how one got there. If z is a function of two variables, x and y, then $z(x,y)$ is a state function, since z is known once x and y are specified. The differential of z is

$$dz = M\,dx + N\,dy$$

The line integral

$$\int_C (M\,dx + N\,dy)$$

is independent of the path in x-y space if and only if

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} \tag{3-42}$$

The total differential can be written as

$$dz = \left(\frac{\partial z}{\partial x}\right)_y dx + \left(\frac{\partial z}{\partial y}\right)_x dy \tag{3-43}$$

and the following condition guarantees path independence.

$$\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right)_y = \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right)_x$$

or

$$\frac{\partial^2 z}{\partial y\,\partial x} = \frac{\partial^2 z}{\partial x\,\partial y} \tag{3-44}$$

Example Suppose z is constant and apply Eq. (3-43).

$$\left[0 = \left(\frac{\partial z}{\partial x}\right)_y dx + \left(\frac{\partial z}{\partial y}\right)_x dy\right]_z$$

Rearrangement gives

$$\left(\frac{\partial z}{\partial x}\right)_y = -\left(\frac{\partial y}{\partial x}\right)_z \left(\frac{\partial z}{\partial y}\right)_x = -\frac{(\partial y/\partial x)_z}{(\partial y/\partial z)_x} \tag{3-45}$$

Alternatively, divide Eq. (3-43) by dy when holding some other variable w constant to obtain

$$\left(\frac{\partial z}{\partial y}\right)_w = \left(\frac{\partial z}{\partial x}\right)_v \left(\frac{\partial x}{\partial y}\right)_w + \left(\frac{\partial z}{\partial y}\right)_x \tag{3-46}$$

Also divide both numerator and denominator of a partial derivative by dw while holding a variable y constant to get

$$\left(\frac{\partial z}{\partial x}\right)_y = \frac{(\partial z/\partial w)_y}{(\partial x/\partial w)_y} = \left(\frac{\partial z}{\partial w}\right)_y \left(\frac{\partial w}{\partial x}\right)_y \tag{3-47}$$

Themodynamic State Functions In thermodynamics, the state functions include the internal energy, U; enthalpy, H; and Helmholtz and Gibbs free energies, A and G, respectively, defined as follows:

$$H = U + p\,V$$
$$A = U - TS$$
$$G = H - TS = U + pV - TS = A + pV$$

S is the entropy, T the absolute temperature, p the pressure, and V the volume. These are also state functions, in that the entropy is specified once two variables (like T and p) are specified, for example. Likewise, V is specified once T and p are specified; it is therefore a state function.

All applications are for closed systems with constant mass. If a process is reversible and only p-V work is done, the first law and differentials can be expressed as follows.

$$dU = T\,dS - p\,dV$$
$$dH = T\,dS + V\,dp$$
$$dA = -S\,dT - p\,dV$$
$$dG = -S\,dT + V\,dp$$

Alternatively, if the internal energy is considered a function of S and V, then the differential is:

$$dU = \left(\frac{\partial U}{\partial S}\right)_V dS + \left(\frac{\partial U}{\partial V}\right)_S dV$$

This is the equivalent of Eq. (3-43) and gives the following definitions.

$$T = \left(\frac{\partial U}{\partial S}\right)_V, \quad p = -\left(\frac{\partial U}{\partial V}\right)_S$$

Since the internal energy is a state function, then Eq. (3-44) must be satisfied.

$$\frac{\partial^2 U}{\partial V\,\partial S} = \frac{\partial^2 U}{\partial S\,\partial V}$$

This is

$$\left(\frac{\partial T}{\partial V}\right)_S = -\left(\frac{\partial p}{\partial S}\right)_V$$

This is one of the Maxwell relations, and the other Maxwell relations can be derived in a similar fashion by applying Eq. (3-44).

$$\left(\frac{\partial T}{\partial p}\right)_S = \left(\frac{\partial V}{\partial S}\right)_p$$

$$\left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial p}{\partial T}\right)_V$$

$$\left(\frac{\partial S}{\partial p}\right)_T = -\left(\frac{\partial V}{\partial T}\right)_p$$

In process simulation it is necessary to calculate enthalpy as a function of state variables. This is done using the following formulas, derived from the above relations by considering S and H as functions of T and p.

$$dH = C_p\,dT + \left[V - T\left(\frac{\partial V}{\partial T}\right)_p\right] dp$$

Enthalpy differences are then given by the following formula.

$$H(T_2, p_2) - H(T_1, p_1) = \int_{T_1}^{T_2} C_p(T, p_1)\,dT + \int_{p_1}^{p_2} \left[V - T\left(\frac{\partial V}{\partial T}\right)_p\right]\Bigg|_{T_2,p} dp$$

The same manipulations can be done for internal energy as a function of T and V.

$$dU = C_V\,dT - \left[p + T\frac{(\partial V/\partial T)_p}{(\partial V/\partial p)_T}\right] dV$$

Partial Derivatives of All Thermodynamic Functions The various partial derivatives of the thermodynamic functions can be classified into six groups. In the general formulas below, the variables U, H, A, G or S are denoted by Greek letters, while the variables V, T, or p are denoted by Latin letters.

Type I (3 possibilities plus reciprocals)

$$\text{General: } \left(\frac{\partial \alpha}{\partial \beta}\right)_\gamma; \text{ Specific: } \left(\frac{\partial p}{\partial T}\right)_V$$

Eq. (3-45) gives

$$\left(\frac{\partial p}{\partial T}\right)_V = -\left(\frac{\partial V}{\partial T}\right)_p \left(\frac{\partial p}{\partial V}\right)_T = -\frac{(\partial V/\partial T)_p}{(\partial V/\partial p)_T}$$

Type II (30 possibilities)

$$\text{General: } \left(\frac{\partial \alpha}{\partial \beta}\right)_\gamma; \text{ Specific: } \left(\frac{\partial G}{\partial T}\right)_V$$

The differential for G gives

$$\left(\frac{\partial G}{\partial T}\right)_V = -S + V\left(\frac{\partial p}{\partial T}\right)_V$$

Using the other equations for U, H, A, or S gives the other possibilities.

Type III (15 possibilities plus reciprocals)

$$\text{General: } \left(\frac{\partial a}{\partial b}\right)_\alpha; \text{ Specific: } \left(\frac{\partial V}{\partial T}\right)_S$$

First expand the derivative using Eq. (3-45).

$$\left(\frac{\partial V}{\partial T}\right)_S = -\left(\frac{\partial S}{\partial T}\right)_V \left(\frac{\partial V}{\partial S}\right)_T = -\frac{(\partial S/\partial T)_V}{(\partial S/\partial V)_T}$$

Then evaluate the numerator and denominator as type II derivatives.

$$\left(\frac{\partial V}{\partial T}\right)_S = -\frac{\dfrac{C_V}{T}}{-\left(\dfrac{\partial V}{\partial T}\right)_p \left(\dfrac{\partial p}{\partial V}\right)_T} = \frac{C_V}{T}\frac{\left(\dfrac{\partial V}{\partial p}\right)_T}{\left(\dfrac{\partial V}{\partial T}\right)_p}$$

These derivatives are of importance for reversible, adiabatic processes (such as in an ideal turbine or compressor), since then the entropy is constant. An example is the Joule-Thomson coefficient.

$$\left(\frac{\partial T}{\partial p}\right)_H = \frac{1}{C_p}\left[-V + T\left(\frac{\partial V}{\partial T}\right)_p\right]$$

Type IV (30 possibilities plus reciprocals)

$$\text{General: } \left(\frac{\partial \alpha}{\partial \beta}\right)_c; \text{ Specific: } \left(\frac{\partial G}{\partial A}\right)_p$$

Use Eq. (3-47) to introduce a new variable.

$$\left(\frac{\partial G}{\partial A}\right)_p = \left(\frac{\partial G}{\partial T}\right)_p \left(\frac{\partial T}{\partial A}\right)_p = \frac{(\partial G/\partial T)_p}{(\partial A/\partial T)_p}$$

This operation has created two type II derivatives; by substitution we obtain

$$\left(\frac{\partial G}{\partial A}\right)_p = \frac{S}{S + p\,(\partial V/\partial T)_p}$$

Type V (60 possibilities)

$$\text{General: } \left(\frac{\partial \alpha}{\partial b}\right)_\beta; \text{ Specific: } \left(\frac{\partial G}{\partial p}\right)_A$$

Start from the differential for dG. Then we get

$$\left(\frac{\partial G}{\partial p}\right)_A = -S\left(\frac{\partial T}{\partial p}\right)_A + V$$

The derivative is type III and can be evaluated by using Eq. (3-45).

$$\left(\frac{\partial G}{\partial p}\right)_A = S\frac{(\partial A/\partial p)_T}{(\partial A/\partial T)_p} + V$$

The two type II derivatives are then evaluated.

$$\left(\frac{\partial G}{\partial p}\right)_A = \frac{Sp\,(\partial V/\partial p)_T}{S + p\,(\partial V/\partial T)_p} + V$$

These derivatives are also of interest for free expansions or isentropic changes.

Type VI (30 possibilities plus reciprocals)

$$\text{General: } \left(\frac{\partial \alpha}{\partial \beta}\right)_\gamma; \text{ Specific: } \left(\frac{\partial G}{\partial A}\right)_H$$

We use Eq. (3-47) to obtain two type V derivatives.

$$\left(\frac{\partial G}{\partial A}\right)_H = \frac{(\partial G/\partial T)_H}{(\partial A/\partial T)_H}$$

These can then be evaluated using the procedures for Type V derivatives.

INTEGRAL CALCULUS

Indefinite Integral If $f'(x)$ is the derivative of $f(x)$, an antiderivative of $f'(x)$ is $f(x)$. Symbolically, the indefinite integral of $f'(x)$ is

$$\int f'(x)\,dx = f(x) + c$$

where c is an arbitrary constant to be determined by the problem. By virtue of the known formulas for differentiation the following relationships hold (a is a constant):

$$\int (du + dv + dw) = \int du + \int dv + \int dw \tag{3-48}$$

$$\int a\,dv = a\int dv \tag{3-49}$$

$$\int v^n\,dv = \frac{v^{n+1}}{n+1} + c \quad (n \neq -1) \tag{3-50}$$

$$\int \frac{dv}{v} = \ln|v| + c \tag{3-51}$$

$$\int a^v\,dv = \frac{a^v}{\ln a} + c \tag{3-52}$$

$$\int e^v\,dv = e^v + c \tag{3-53}$$

$$\int \sin v\,dv = -\cos v + c \tag{3-54}$$

$$\int \cos v\,dv = \sin v + c \tag{3-55}$$

$$\int \sec^2 v\,dv = \tan v + c \tag{3-56}$$

$$\int \csc^2 v\,dv = -\cot v + c \tag{3-57}$$

$$\int \sec v \tan v\,dv = \sec v + c \tag{3-58}$$

$$\int \csc v \cot v\,dv = -\csc v + c \tag{3-59}$$

$$\int \frac{dv}{v^2 + a^2} = \frac{1}{a}\tan^{-1}\frac{v}{a} + c \tag{3-60}$$

$$\int \frac{dv}{\sqrt{a^2 - v^2}} = \sin^{-1}\frac{v}{a} + c \tag{3-61}$$

$$\int \frac{dv}{v^2 - a^2} = \frac{1}{2a}\ln\left|\frac{v-a}{v+a}\right| + c \tag{3-62}$$

$$\int \frac{dv}{\sqrt{v^2 \pm a^2}} = \ln|v + \sqrt{v^2 \pm a^2}| + c \tag{3-63}$$

$$\int \sec v\,dv = \ln(\sec v + \tan v) + c \tag{3-64}$$

$$\int \csc v\,dv = \ln(\csc v - \cot v) + c \tag{3-65}$$

Example Derive $\int a^v\,dv = (a^v/\ln a) + c$. By reference to the differentiation formula $da^v/dv = a^v \ln a$, or in the more usable form $d(a^v/\ln a) = a^v\,dv$, let $f' = a^v\,dv$; then $f = a^v/\ln a$ and hence $\int a^v\,dv = (a^v/\ln a) + c$.

Example Find $\int (3x^2 + e^x - 10)\,dx$ using Eq. (3-48). $\int (3x^2 + e^x - 10)\,dx = 3\int x^2\,dx + \int e^x\,dx - 10\int dx = x^3 + e^x - 10x + c$ (by Eqs. 3-50, 3-53).

Example Find $\int \dfrac{7x\,dx}{2 - 3x^2}$. Let $v = 2 - 3x^2$; $dv = -6x\,dx$

Thus

$$\int \frac{7x\,dx}{2-3x^2} = 7\int \frac{x\,dx}{2-3x^2} = -\frac{7}{6}\int \frac{-6x\,dx}{2-3x^2}$$

$$= -\frac{7}{6}\int \frac{dv}{v}$$

$$= -\frac{7}{6}\ln|v| + c$$

$$= -\frac{7}{6}\ln|2 - 3x^2| + c$$

Example—Constant of Integration By definition the derivative of x^3 is $3x^2$, and x^3 is therefore the integral of $3x^2$. However, if $f = x^3 + 10$, it follows that $f' = 3x^2$, and $x^3 + 10$ is therefore also the integral of $3x^2$. For this reason the constant c in $\int 3x^2\,dx = x^3 + c$ must be determined by the problem conditions, i.e., the value of f for a specified x.

Methods of Integration In practice it is rare when generally encountered functions can be directly integrated. For example, the integrand in $\int \sqrt{\sin x}\,dx$ which appears quite simple has no elementary function whose derivative is $\sqrt{\sin x}$. In general, there is no explicit way of determining whether a particular function can be integrated into an elementary form. As a whole, integration is a trial-and-error proposition which depends on the effort and ingenuity of the practitioner. The following are general procedures which can be used to find the elementary forms of the integral when they exist. When they do not exist or cannot be found either from tabled integration formulas or directly, the only recourse is series expansion as illustrated later. Indefinite integrals cannot be solved numerically unless they are redefined as definite integrals (see "Definite Integral"), i.e., $F(x) = \int f(x)\,dx$, indefinite, whereas $F(x) = \int_a^x f(t)\,dt$, definite.

Direct Formula Many integrals can be solved by transformation in the integrand to one of the forms given previously.

Example Find $\int x^2 \sqrt{3x^3 + 10}\,dx$. Let $v = 3x^3 + 10$ for which $dv = 9x^2\,dx$. Thus

$$\int x^2 \sqrt{3x^3 + 10}\,dx = \int (3x^3 + 10)^{1/2}\,(x^2\,dx)$$

$$= \frac{1}{9}\int (3x^3 + 10)^{1/2}(9x^2\,dx)$$

$$= \frac{1}{9}\int v^{1/2}\,dv$$

$$= \frac{1}{9}\,\frac{v^{3/2}}{3/2} + c \qquad \text{[by Eq. (3-50)]}$$

$$= \frac{2}{27}(3x^3 + 10)^{3/2} + c$$

Trigonometric Substitution This technique is particularly well adapted to integrands in the form of radicals. For these the function is transformed into a trigonometric form. In the latter form they may be more easily recognizable relative to the identity formulas. These functions and their transformations are

$$\sqrt{x^2 - a^2} \quad \text{Let } x = a \sec \theta$$
$$\sqrt{x^2 + a^2} \quad \text{Let } x = a \tan \theta$$
$$\sqrt{a^2 - x^2} \quad \text{Let } x = a \sin \theta$$

Example Find $\int \frac{\sqrt{4 - 9x^2}}{x^2}\,dx$. Let $x = \frac{2}{3}\sin\theta$; then $dx = \frac{2}{3}\cos\theta\,d\theta$.

$$3\int \frac{\sqrt{(2/3)^2 - x^2}}{x^2}\,dx = 3\int \frac{2/3\sqrt{1 - \sin^2\theta}}{(2/3)^2\sin^2\theta}\left(\frac{2}{3}\cos\theta\,d\theta\right)$$

$$= 3\int \frac{\cos^2\theta}{\sin^2\theta}\,d\theta$$

$$= 3\int \cot^2\theta\,d\theta$$

$$= -3\cot\theta - 3\theta + c \text{ by trigonometric transform}$$

$$= -\frac{\sqrt{4 - 9x^2}}{x} - 3\sin^{-1}\frac{3}{2}x + c \text{ in terms of } x$$

Algebraic Substitution Functions containing elements of the type $(a + bx)^{1/n}$ are best handled by the algebraic transformation $y^n = a + bx$.

Example Find $\int \frac{x\,dx}{(3 + 4x)^{1/4}}$. Let $3 + 4x = y^4$; then $4dx = 4y^3\,dy$ and

$$\int \frac{x\,dx}{(3 + 4x)^{1/4}} = \int \frac{\frac{y^4 - 3}{4}\,y^3\,dy}{y}$$

$$= \frac{1}{4}\int y^2(y^4 - 3)\,dy$$

$$= \frac{1}{4}\,\frac{y^7}{7} - \frac{3}{4}\,\frac{y^3}{3} + c$$

$$= \frac{1}{28}(3 + 4x)^{7/4} - \frac{1}{4}(3 + 4x)^{3/4} + c$$

General The number of possible transformations one might use are unlimited. No specific overall rules can be given. Success in handling integration problems depends primarily upon experience and ingenuity. The following example illustrates the extent to which alternative approaches are possible.

Example Find $\int \frac{dx}{e^x - 1}$. Let $e^x = y$; then $e^x\,dx = dy$ or $dx = 1/y\,dy$.

$$\int \frac{dx}{e^x - 1} = \int \frac{(1/y)\,dy}{y - 1} = \int \frac{dy}{y^2 - y} = \ln\frac{y - 1}{y} = \ln\frac{e^x - 2}{e^x}$$

Partial Fractions Rational functions are of the type $f(x)/g(x)$ where $f(x)$ and $g(x)$ are polynomial expressions of degrees m and n respectively. If the degree of f is higher than g, perform the algebraic division—the remainder will then be at least one degree less than the denominator. Consider the following types:

Type 1 Reducible denominator to linear unequal factors. For example,

$$\frac{1}{x^3 - x^2 - 4x + 4} = \frac{1}{(x + 2)(x - 2)(x - 1)}$$

$$= \frac{A}{x + 2} + \frac{B}{x - 2} + \frac{C}{x - 1}$$

$$= \frac{A(x - 2)(x - 1) + B(x + 2)(x - 1) + C(x + 2)(x - 2)}{(x + 2)(x - 2)(x - 1)}$$

$$= \frac{x^2(A + B + C) + x(-3A + B) + (2A - 2B - 4C)}{(x + 2)(x - 2)(x - 1)}$$

Equate coefficients and solve for A, B, and C.

$$A + B + C = 0$$
$$-3A + B = 0$$
$$2A - 2B - 4C = 1$$
$$A = 1/12,\ B = 1/4,\ C = -1/3$$

$$\frac{1}{x^3 - x^2 - 4x + 4} = \frac{1}{12(x + 2)} + \frac{1}{4(x - 2)} - \frac{1}{3(x - 1)}$$

Hence

$$\int \frac{dx}{x^3 - x^2 - 4x + 4} = \int \frac{dx}{12(x + 2)} + \int \frac{dx}{4(x - 2)} - \int \frac{dx}{3(x - 1)}$$

Parts An extremely useful formula for integration is the relation

$$d(uv) = u\,dv + v\,du$$

and

$$uv = \int u\,dv + \int v\,du$$

or

$$\int u\,dv = uv - \int v\,du$$

No general rule for breaking an integrand can be given. Experience alone limits the use of this technique. It is particularly useful for trigonometric and exponential functions.

Example Find $\int xe^x\,dx$. Let

$$u = x \quad \text{and} \quad dv = e^x\,dx$$
$$du = dx \qquad v = e^x$$

Therefore
$$\int xe^x\,dx = xe^x - \int e^x\,dx$$
$$= xe^x - e^x + c$$

Example Find $\int e^x \sin x\,dx$. Let
$$u = e^x \qquad dv = \sin x\,dx$$
$$du = e^x\,dx \qquad v = -\cos x$$
$$\int e^x \sin x\,dx = -e^x \cos x + \int e^x \cos x\,dx$$

Again
$$u = e^x \qquad dv = \cos x\,dx$$
$$du = e^x\,dx \qquad v = \sin x$$
$$\int e^x \sin x\,dx = -e^x \cos x + e^x \sin x - \int e^x \sin x\,dx + c$$
$$= (e^x/2)(\sin x - \cos x) + \frac{c}{2}$$

Series Expansion When an explicit function cannot be found, the integration can sometimes be carried out by a series expansion.

Example Find $\int e^{-x^2}\,dx$. Since
$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \cdots$$
$$\int e^{-x^2}\,dx = \int dx - \int x^2\,dx + \int \frac{x^4}{2!}\,dx - \int \frac{x^6}{3!}\,dx + \cdots$$
$$= x - \frac{x^3}{3} + \frac{x^5}{5.2!} - \frac{x^7}{7.3!} + \cdots \qquad \text{for all } x$$

Definite Integral The concept and derivation of the definite integral are completely different from those for the indefinite integral. These are by definition different types of operations. However, the formal operation $\int$ as it turns out treats the integrand in the same way for both.

Consider the function $f(x) = 10 - 10e^{-2x}$. Define $x_1 = a$ and $x_n = b$, and suppose it is desirable to compute the area between the curve and the coordinate axis $y = 0$ and bounded by $x_1 = a$, $x_n = b$. Obviously, by a sufficiently large number of rectangles this area could be approximated as closely as desired by the formula
$$\sum_{i=1}^{n-1} f(\xi_i)(x_{i+1} - x_i) = f(\xi_1)(x_2 - a) + f(\xi_2)(x_3 - x_2) + \cdots + f(\xi_{n-1})(b - x_{n-1}) \qquad x_{i-1} \le \xi_{i-1} \le x_i$$

The definite integral of $f(x)$ is defined as
$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(\xi_i)(x_{i+1} - x_i)$$

where the points $x_1, x_2, \ldots, x_n$ are equally spaced. For a rigorous definition of the definite integral the references should be consulted.

Thus, the value of a definite integral depends on the limits a, b, and any selected variable coefficients in the function but not on the dummy variable of integration x. Symbolically
$$F(x) = \int f(x)\,dx \qquad \text{indefinite integral where } dF/dx = f(x)$$
or
$$F(a, b) = \int_a^b f(x)\,dx \qquad \text{definite integral}$$
$$F(\alpha) = \int_a^b f(x, \alpha)\,dx$$

There are certain restrictions of the integration definition, "The function $f(x)$ must be continuous in the finite interval (a, b) with at most a finite number of finite discontinuities," which must be observed before integration formulas can be generally applied. Two of these restrictions give rise to so-called **improper integrals** and require special handling. These occur when

1. The limits of integration are not both finite, i.e., $\int_0^\infty e^{-x}\,dx$.
2. The function becomes infinite within the interval of integration, i.e.,
$$\int_0^1 \frac{1}{\sqrt{x}}\,dx$$

Techniques for determining when integration is valid under these conditions are available in the references. However, the following simplified rules will, in general, serve as a guide for most practical applications.

Rule 1 For the integral
$$\int_0^\infty \frac{\phi(x)}{x^n}\,dx$$
if $\phi(x)$ is bounded, the integral will converge for $n > 1$ and not converge for $n \le 1$.

It is easily seen that $\int_0^\infty e^{-x}\,dx$ converges by noting $1/x^2 > 1/e^x > 0$ for large x.

Rule 2 For the integral
$$\int_a^b \frac{\phi(x)}{(a - x)^n}\,dx,$$
if $\phi(x)$ is bounded, the integral will converge for $n < 1$ and diverge for $n \ge 1$. Thus
$$\int_0^1 \frac{1}{\sqrt{x}}\,dx$$
will converge (exist) since $\frac{1}{2} = n < 1$.

Properties The fundamental theorem of calculus states
$$\int_a^b f(x)\,dx = F(b) - F(a)$$
where
$$dF(x)/dx = f(x)$$

Other properties of the definite integral are
$$\int_a^b c[f(x)\,dx] = c\int_a^b f(x)\,dx$$
$$\int_a^b [f_1(x) + f_2(x)]\,dx = \int_a^b f_1(x)\,dx + \int_a^b f_2(x)\,dx$$
$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx$$
$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$
$$\int_a^b f(x)\,dx = (b - a)f(\xi) \text{ for some } \xi \text{ in } (a, b)$$
$$\frac{\partial}{\partial b}\int_a^b f(x)\,dx = f(b)$$
$$\frac{\partial}{\partial a}\int_a^b f(x)\,dx = -f(a)$$
$$\frac{dF(\alpha)}{d\alpha} = \int_a^b \frac{\partial f(x, \alpha)}{\partial \alpha}\,dx \text{ if } a \text{ and } b \text{ are constant}$$
$$\int_a^b dx \int_c^d f(x, \alpha)\,d\alpha = \int_c^d d\alpha \int_a^b f(x, \alpha)\,dx \qquad (3\text{-}66)$$

when $F(x) = \int_{a(x)}^{b(x)} f(x, y)\,dy$

the Leibniz rule gives
$$\frac{dF}{dx} = \frac{db}{dx} f[x, b(x)] - \frac{da}{dx} f[x, a(x)] + \int_{a(x)}^{b(x)} \frac{\partial f}{\partial x}\,dy$$

Example Find $\int_0^{\pi/2} \sin x\,dx$.
$$\int_0^{\pi/2} \sin x\,dx = [-\cos x]_0^{\pi/2} = -\left(\cos\frac{\pi}{2} - \cos 0\right) = 1$$

since
$$-d\cos x/dx = \sin x$$

Example Find $\int_0^2 \frac{dx}{(x-1)^2}$. Direct application of the formula would yield the incorrect value

$$\int_0^2 \frac{dx}{(x-1)^2} = \left[-\frac{1}{x-1}\right]_0^2 = -2$$

It should be noted that $f(x) = 1/(x-1)^2$ becomes unbounded as $x \to 1$ and by Rule 2 the integral diverges and hence is said not to exist.

Methods of Integration All the methods of integration available for the indefinite integral can be used for definite integrals. In addition, several others are available for the latter integrals and are indicated below.

Change of Variable This substitution is basically the same as previously indicated for indefinite integrals. However, for definite integrals, the limits of integration must also be changed: i.e., for $x = \phi(t)$,

$$\int_a^b f(x)\,dx = \int_{t_0}^{t_1} f[\phi(t)]\phi'(t)\,dt$$

where $t = t_0$ when $x = a$
$t = t_1$ when $x = b$

Example Find $\int_0^4 \sqrt{16 - x^2}\,dx$. Let

$$x = 4\sin\theta \qquad (x = 0,\ \theta = 0)$$
$$dx = 4\cos\theta\,d\theta \quad (x = 4,\ \theta = \pi/2)$$

Then $$\int_0^4 \sqrt{16 - x^2}\,dx = 16\int_0^{\pi/2} \cos^2\theta\,d\theta = 16[\tfrac{1}{2}\theta + \tfrac{1}{4}\sin 2\theta]_0^{\pi/2} = 4\pi$$

Differentiation Here the application of the general rules for differentiating under the integral sign may be useful.

Example Find

$$\phi(\alpha) = \int_0^\infty \frac{e^{-\alpha x}\sin x}{x}\,dx \quad (\alpha > 0)$$

Since this is a continuous function of α, it may be differentiated under the integral sign

$$\frac{d\phi}{d\alpha} = -\int_0^\infty e^{-\alpha x}\sin x\,dx$$
$$= -1/(1+\alpha^2)$$
$$\phi(\alpha) = -\tan^{-1}\alpha + c$$

and since $\phi(\alpha) \to 0$ as $\alpha \to \infty$,

$$c = \pi/2$$
$$\phi(\alpha) = -\tan^{-1}\alpha + \pi/2$$

Integration It is sometimes useful to generate a double integral to solve a problem. By this approach, the fundamental theorem indicated by Eq. (3-66) can be used.

Example Find $\int_0^1 \frac{x^b - x^\alpha}{\ln x}\,dx$

Consider $$\int_0^1 x^\alpha\,dx = \frac{1}{\alpha+1} \quad (\alpha > -1)$$

Then multiplying both sides by $d\alpha$ and integrating between a and b,

$$\int_a^b d\alpha \int_0^1 x^\alpha\,dx = \int_a^b \frac{d\alpha}{\alpha+1} = \ln\left|\frac{b+1}{a+1}\right|$$

But also

$$\int_a^b d\alpha \int_0^1 x^\alpha\,dx = \int_0^1 dx \int_a^b x^\alpha\,d\alpha = \int_0^1 \frac{x^b - x^\alpha}{\ln x}\,dx$$

Therefore $$\int_0^1 \frac{x^b - x^\alpha}{\ln x}\,dx = \ln\left|\frac{b+1}{a+1}\right|$$

Complex Variable Certain definite integrals can be evaluated by the technique of complex variable integration. This is described in the references for "Complex Variables."

Numerical Because of the property of definite integrals another method for obtaining their solution is available which cannot be applied to indefinite integrals. This involves a numerical approximation based on the previously outlined summation definition:

$$\lim_{n\to\infty} \sum_1^{n-1} f(\xi_i)(x_{i+1} - x_i) = \int_a^b f(x)\,dx$$

where $x_1 = a$ and $x_n = b$

Examples of this procedure are given in the subsection "Numerical Analysis and Approximate Methods."

INFINITE SERIES

REFERENCES: 53, 126, 127, 163. For asymptotic series and asymptotic methods, see Refs. 51, 127.

DEFINITIONS

A succession of numbers or terms that are formed according to some definite rule is called a sequence. The indicated sum of the terms of a sequence is called a series. A series of the form $a_0 + a_1(x-c) + a_2(x-c)^2 + \cdots + a_n(x-c)^n + \cdots$ is called a power series.

Consider the sum of a finite number of terms in the geometric series (a special case of a power series).

$$S_n = a + ar + ar^2 + ar^3 + \cdots + ar^{n-1} \tag{3-67}$$

For any number of terms n, the sum equals

$$S_n = a\,\frac{1-r^n}{1-r}$$

In this form, the geometric series is assumed finite.

In the form of Eq. (3-67), it can further be defined that the terms in the series be nonending and therefore an infinite series.

$$S = a + ar + ar^2 + \cdots + ar^n + \cdots \tag{3-68}$$

However, the defined sum of the terms [Eq. (3-67)]

$$S_n = a\,\frac{1-r^n}{1-r} \qquad r \neq 1$$

while valid for any finite value of r and n now takes on a different interpretation. In this sense it is necessary to consider the limit of S_n as n increases indefinitely:

$$S = \lim_{n\to\infty} S_n$$
$$= a \lim_{n\to\infty} \frac{1-r^n}{1-r}$$

For this, it is stated the infinite series converges if the limit of S_n approaches a fixed finite value as n approaches infinity. Otherwise, the series is **divergent.**

On this basis an analysis of

$$S = a \lim_{n\to\infty} \frac{1-r^n}{1-r}$$

shows that if r is less than 1 but greater than -1, the infinite series is convergent. For values outside of the range $-1 < r < 1$, the series is divergent because the sum is not defined. The range $-1 < r < 1$ is called the **region of convergence.** (We assume $a \neq 0$.)

Consider the divergence of Eq. (3-68) when $r = -1$ and $+1$. For the former case $r = -1$,

$$S = a + a(-1) + a(-1)^2 + a(-1)^3 + \cdots + a(-1)^n + \cdots$$
$$= a - a + a - a + a - \cdots$$

and for which

$$S = a \lim_{n\to\infty} \frac{1-r^n}{1-r}$$

$$= a \lim_{n\to\infty} \frac{1-(-1)^n}{1+1} \quad \text{undefined limit (if } a \neq 0)$$

Since the limit sum does not exist, the series is divergent. This is defined as a bounded or oscillating divergent series. Similarly for the value $r = +1$,

$$S = a + a(1) + a(1)^2 + a(1)^3 + \cdots + a(1)^n + \cdots$$

$$S = a + a + a + a + \cdots + a + \cdots \qquad (a \neq 0)$$

The series is also divergent but defined as an **unbounded divergent series.**

There are also two types of convergent series. Consider the new series

$$S = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + (-1)^{n+1}\frac{1}{n} + \cdots \tag{3-69}$$

It can be shown that the series (3-69) does converge to the value $S = \log 2$. However, if each term is replaced by its absolute value, the series becomes unbounded and therefore divergent (unbounded divergent):

$$S = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots \tag{3-70}$$

In this case the series (3-69) is defined as a conditionally convergent series. If the replacement series of absolute values also converges, the series is defined to converge absolutely.

Series (3-69) is further defined as an alternating series, while series (3-70) is referred to as a positive series.

OPERATIONS WITH INFINITE SERIES

1. The convergence or divergence of an infinite series is unaffected by the removal of a finite number of finite terms. This is a trivial theorem but useful to remember, especially when using the comparison test to be described in the subsection "Tests for Convergence and Divergence."

2. If a series is conditionally convergent, its sums can be made to have any arbitrary value by a suitable rearrangement of the series; it can in fact be made divergent or oscillatory (Riemann's theorem). This seemingly paradoxical theorem can be illustrated by the following example.

Example $S = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \cdots$

The series is rearranged so that each positive term is followed by two negative terms:

$$t = 1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \frac{1}{5} - \frac{1}{10} - \frac{1}{12} + \cdots$$

Define t_{3n} for the first $3n$ terms in the series

$$t_{3n} = \left(1 - \frac{1}{2}\right) - \frac{1}{4} + \left(\frac{1}{3} - \frac{1}{6}\right) - \frac{1}{8} + \cdots + \left(\frac{1}{2n-1} - \frac{1}{4n-2}\right) - \frac{1}{4n}$$

$$= \frac{1}{2} - \frac{1}{4} + \frac{1}{6} - \frac{1}{8} + \cdots + \frac{1}{4n-2} - \frac{1}{4n}$$

$$= \frac{1}{2}\left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{1}{2n-1} - \frac{1}{2n}\right)$$

$$= \frac{1}{2} S_{2n}$$

where S_{2n} is the sum of the first $2n$ terms of the original series. Thus

$$\lim_{n\to\infty} t_{3n} = \lim_{n\to\infty} \frac{1}{2} S_{2n}$$

$$t = \frac{1}{2} S$$

and since $\lim t_{3n+2} = \lim t_{3n+1} = \lim t_{3n}$, it follows the sum of the series t is (½) S. Hence a rearrangement of the terms of an alternating series alters the sum of the series.

3. A series of positive terms, if convergent, has a sum independent of the order of its terms; but if divergent, it remains divergent however its terms are rearranged.

4. An oscillatory series can always be made to converge by grouping the terms in brackets.

Example Consider the series

$$1 - \frac{1}{2} + \frac{2}{3} - \frac{3}{4} + \frac{4}{5} - \frac{5}{6} + \cdots$$

which oscillates between the values 0.306 and 1.306. However, the series

$$\left(1 - \frac{1}{2}\right) + \left(\frac{2}{3} - \frac{3}{4}\right) + \left(\frac{4}{5} - \frac{5}{6}\right) + \cdots = \frac{1}{2} - \frac{1}{12} - \frac{1}{30} - \frac{1}{56} - \cdots \cong 0.306\cdots$$

and

$$1 - \left(\frac{1}{2} - \frac{2}{3}\right) - \left(\frac{3}{4} - \frac{4}{5}\right) - \left(\frac{5}{6} - \frac{6}{7}\right) + \cdots = 1 + \frac{1}{6} + \frac{1}{20} + \frac{1}{42} + \cdots = 1.306\cdots$$

5. A power series can be inverted, provided the first-degree term is not zero. Given

$$y = b_1x + b_2x^2 + b_3x^3 + b_4x^4 + b_5x^5 + b_6x^6 + b_7x^7 + \cdots$$

then $$x = B_1y + B_2y^2 + B_3y^3 + B_4y^4 + B_5y^5 + B_6y^6 + B_7y^7 + \cdots$$

where $B_1 = 1/b_1$

$B_2 = -b_2/b_1^3$

$B_3 = (1/b_1^5)(2b_2^2 - b_1b_3)$

$B_4 = (1/b_1^7)(5b_1b_2b_3 - b_1^2b_4 - 5b_2^3)$

Additional coefficients are available in the references.

6. Two series may be added or subtracted term by term provided each is a convergent series. The joint sum is equal to the sum (or difference) of the individuals.

7. The sum of two divergent series can be convergent. Similarly, the sum of a convergent series and a divergent series must be divergent.

Example Given

$$\sum_{n=1}^{\infty}\left(\frac{1+n}{n^2}\right) = \frac{2}{1} + \frac{3}{4} + \frac{4}{9} + \frac{5}{16} + \cdots \qquad \text{(a divergent series)}$$

$$\sum_{n=1}^{\infty}\left(\frac{1-n}{n^2}\right) = -\frac{1}{4} - \frac{2}{9} - \frac{3}{16} + \cdots \qquad \text{(a divergent series)}$$

However, $$\sum\left(\frac{1+n}{n^2}\right) + \sum\left(\frac{1-n}{n^2}\right) = \sum\left(\frac{1+n+1-n}{n^2}\right)$$

$$= 2\sum\frac{1}{n^2} \quad \text{(convergent)}$$

8. A power series may be integrated term by term to represent the integral of the function within an interval of the region of convergence. If $f(x) = a_0 + a_1x + a_2x^2 + \cdots$, then

$$\int_{x_1}^{x_2} f(x)\,dx = \int_{x_1}^{x_2} a_0\,dx + \int_{x_1}^{x_2} a_1x\,dx + \int_{x_1}^{x_2} a_2x^2\,dx + \cdots$$

9. A power series may be differentiated term by term and represents the function $df(x)/dx$ within the same region of convergence as $f(x)$.

TESTS FOR CONVERGENCE AND DIVERGENCE

In general, the problem of determining whether a given series will converge or not can require a great deal of ingenuity and resourcefulness. There is no all-inclusive test which can be applied to all series. As the only alternative, it is necessary to apply one or more of the developed theorems in an attempt to ascertain the convergence or divergence of the series under study. The following defined tests are given in relative order of effectiveness. For examples, see references on advanced calculus.

1. *Comparison Test.* A series will converge if the absolute value of each term (with or without a finite number of terms) is less than the corresponding term of a known convergent series. Similarly, a positive series is divergent if it is termwise larger than a known divergent series of positive terms.

2. *nth-Term Test.* A series is divergent if the nth term of the series does not approach zero as n becomes increasingly large.
3. *Ratio Test.* If the absolute ratio of the $(n+1)$ term divided by the nth term as n becomes unbounded approaches
   a. A number less than 1, the series is absolutely convergent
   b. A number greater than 1, the series is divergent
   c. A number equal to 1, the test is inconclusive
4. *Alternating-Series Leibniz Test.* If the terms of a series are alternately positive and negative and never increase in value, the absolute series will converge, provided that the terms tend to zero as a limit.
5. *Cauchy's Root Test.* If the nth root of the absolute value of the nth term, as n becomes unbounded, approaches
   a. A number less than 1, the series is absolutely convergent
   b. A number greater than 1, the series is divergent
   c. A number equal to 1, the test is inconclusive
6. *Maclaurin's Integral Test.* Suppose $\sum a_n$ is a series of positive terms and f is a continuous decreasing function such that $f(x) \geq 0$ for $1 \leq x < \infty$ and $f(n) = a_n$. Then the series and the improper integral $\int_1^\infty f(x)\,dx$ either both converge or both diverge.

SERIES SUMMATION AND IDENTITIES

Sums for the First *n* Numbers to Integer Powers

$$\sum_{j=1}^{n} j = \frac{n(n+1)}{2} = 1 + 2 + 3 + 4 + \cdots + n$$

$$\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6} = 1^2 + 2^2 + 3^2 + 4^2 + \cdots + n^2$$

$$\sum_{j=1}^{n} j^3 = \frac{n^2(n+1)^2}{4} = 1^3 + 2^3 + 3^3 + \cdots + n^3$$

$$\sum_{j=1}^{n} j^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30} = 1^4 + 2^4 + 3^4 + \cdots + n^4$$

Arithmetic Progression

$$\sum_{k=1}^{n} [a + (k-1)d] = a + (a+d) + (a+2d) + (a+3d) + \cdots + [a + (n-1)]d$$
$$= na + \frac{1}{2}n(n-1)d$$

Geometric Progression

$$\sum_{j=1}^{n} ar^{j-1} = a + ar + ar^2 + ar^3 + \cdots + ar^{n-1}$$
$$= a\,\frac{1-r^n}{1-r} \qquad r \neq 1$$

Harmonic Progression

$$\sum_{k=0}^{n} \frac{1}{a+kd} = \frac{1}{a} + \frac{1}{a+d} + \frac{1}{a+2d} + \frac{1}{a+3d} + \frac{1}{a+4d} + \cdots + \frac{1}{a+nd}$$

The reciprocals of the terms of the arithmetic-progression series are called harmonic progression. No general summation formulas are available for this series.

Binomial Series

$$(x+y)^n = x^n + nx^{n-1}y + \frac{n(n-1)}{2!}x^{n-2}y^2 + \frac{n(n-1)(n-2)}{3!}x^{n-3}y^3 + \cdots + \frac{n!}{(n-r)!r!}x^{n-r}y^r + \cdots + y^n$$

$$(1 \pm x)^n = 1 \pm nx + \frac{n(n-1)}{2!}x^2 \pm \frac{n(n-1)(n-2)}{3!}x^3 + \cdots \quad (x^2 < 1)$$

Taylor's Series

$$f(x+h) = f(h) + xf'(h) + \frac{x^2}{2!}f''(h) + \frac{x^3}{3!}f'''(h) + \cdots$$

or $$f(x) = f(x_0) + f'(x_0)\,(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \cdots$$

Example Find a series expansion for $f(x) = \ln(1+x)$ about $x_0 = 0$.

$$f'(x) = (1+x)^{-1}, \quad f''(x) = -(1+x)^{-2}, \quad f'''(x) = 2(1+x)^{-3}, \text{ etc.}$$

thus $$f(0) = 0, \quad f'(0) = 1, \quad f''(0) = -1, \quad f'''(1) = 2, \text{ etc.}$$

$$\ln(x+1) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots + (-1)^{n+1}\frac{x^n}{n} + \cdots$$

which converges for $-1 < x \leq 1$.

Maclaurin's Series

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \frac{x^3}{3!}f'''(0) + \cdots$$

This is simply a special case of Taylor's series when h is set to zero.

Exponential Series

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots \quad -\infty < x < \infty$$

Logarithmic Series

$$\ln x = \frac{x-1}{x} + \frac{1}{2}\left(\frac{x-1}{x}\right)^2 + \frac{1}{3}\left(\frac{x-1}{x}\right)^3 + \cdots \qquad (x > \tfrac{1}{2})$$

$$\ln x = 2\left[\left(\frac{x-1}{x+1}\right) + \frac{1}{3}\left(\frac{x-1}{x+1}\right)^3 + \cdots\right] \quad (x > 0)$$

Trigonometric Series*

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \qquad -\infty < x < \infty$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \qquad -\infty < x < \infty$$

$$\sin^{-1} x = x + \frac{x^3}{6} + \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{x^5}{5} + \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{5}{6}\cdot\frac{x^7}{7} + \cdots \qquad (x^2 < 1)$$

$$\tan^{-1} x = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \cdots \qquad (x^2 < 1)$$

Taylor Series The Taylor series for a function of two variables, expanded about the point (x_0, y_0), is

$$f(x,y) = f(x_0,y_0) + \left.\frac{\partial f}{\partial x}\right|_{x_0,y_0}(x - x_0) + \left.\frac{\partial f}{\partial y}\right|_{x_0,y_0}(y - y_0) + \frac{1}{2!}\left[\left.\frac{\partial^2 f}{\partial x^2}\right|_{x_0,y_0}(x-x_0)^2 + 2\left.\frac{\partial^2 f}{\partial x \partial y}\right|_{x_0,y_0}(x-x_0)(y-y_0) + \left.\frac{\partial^2 f}{\partial y^2}\right|_{x_0,y_0}(y-y_0)^2\right] + \cdots$$

Partial Sums of Infinite Series, and How They Grow Calculus textbooks devote much space to tests for convergence and divergence of series that are of little practical value, since a convergent

* tan x series has awkward coefficients and should be computed as $\left[(\text{sign})\,\dfrac{\sin x}{\sqrt{1-\sin^2 x}}\right]$.

series either converges rapidly, in which case almost any test (among those presented in the preceding subsections) will do; or it converges slowly, in which case it is not going to be of much use unless there is some way to get at its sum without adding up an unreasonable number of terms. To find out, as accurately as possible, how fast a convergent series converges and how fast a divergent series diverges, see Ref. 34.

COMPLEX VARIABLES

REFERENCES: *General.* 73, 163, 172, 179. *Applied and computational complex analysis.* 141, 146, 179.

Numbers of the form $z = x + iy$, where x and y are real, $i^2 = -1$, are called complex numbers. The numbers $z = x + iy$ are representable in the plane as shown in Fig. 3-46. The following definitions and terminology are used:

1. Distance $OP = r$ = modulus of z written $|z|$. $|z| = \sqrt{x^2 + y^2}$.
2. x is the real part of z.
3. y is the imaginary part of z.
4. The angle θ, $0 \le \theta < 2\pi$, measured counterclockwise from the positive x axis to OP is the argument of z. $\theta = \arctan y/x = \arcsin y/r = \arccos x/r$ if $x \ne 0$, $\theta = \pi/2$ if $x = 0$ and $y > 0$.
5. The numbers r, θ are the polar coordinates of z.
6. $z = x - iy$ is the complex conjugate of z.

ALGEBRA

Let $z_1 = x_1 + iy_1$, $z_2 = x_2 + iy_2$.

Equality $z_1 = z_2$ if and only if $x_1 = x_2$ and $y_1 = y_2$.

Addition $z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2)$.

Subtraction $z_1 - z_2 = (x_1 - x_2) + i(y_1 - y_2)$.

Multiplication $z_1 \cdot z_2 = (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1)$.

Division $$z_1/z_2 = \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + i\frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2},\ z_2 \ne 0.$$

SPECIAL OPERATIONS

$z\bar{z} = x^2 + y^2 = |z|^2$; $\overline{z_1 \pm z_2} = \bar{z}_1 \pm \bar{z}_2$; $\bar{\bar{z}}_1 = z_1$; $\overline{z_1z_2} = \bar{z}_1\bar{z}_2$; $|z_1 \cdot z_2| = |z_1| \cdot |z_2|$; $\arg(z_1 \cdot z_2) = \arg z_1 + \arg z_2$; $\arg(z_1/z_2) = \arg z_1 - \arg z_2$; $i^{4n} = 1$ for n any integer; $i^{2n} = -1$ where n is any odd integer; $z + \bar{z} = 2x$; $z - z = 2iy$.

Every complex quantity can be expressed in the form $x + iy$.

TRIGONOMETRIC REPRESENTATION

By referring to Fig. 3-46, there results $x = r\cos\theta$, $y = r\sin\theta$ so that $z = x + iy = r(\cos\theta + i\sin\theta)$, which is called the polar form of the complex number. $\cos\theta + i\sin\theta = e^{i\theta}$. Hence $z = x + iy = re^{i\theta}$. $\bar{z} = x - iy = re^{-i\theta}$. Two important results from this are $\cos\theta = (e^{i\theta} + e^{-i\theta})/2$ and $\sin\theta = (e^{i\theta} - e^{-i\theta})/2i$. Let $z_1 = r_1e^{i\theta_1}$, $z_2 = r_2e^{i\theta_2}$. This form is convenient for multiplication for $z_1z_2 = r_1r_2e^{i(\theta_1 + \theta_2)}$ and for division for $z_1/z_2 = (r_1/r_2)e^{i(\theta_1 - \theta_2)}$, $z_2 \ne 0$.

POWERS AND ROOTS

If n is a positive integer, $z^n = (re^{i\theta})^n = r^ne^{in\theta} = r^n(\cos n\theta + i\sin n\theta)$.

If n is a positive integer,

$$z^{1/n} = r^{1/n}e^{i[(\theta + 2k\pi)/n]} = r^{1/n}\left[\cos\left(\frac{\theta + 2k\pi}{n}\right) + i\sin\left(\frac{\theta + 2k\pi}{n}\right)\right]$$

and selecting values of $k = 0, 1, 2, 3, \ldots, n - 1$ give the n distinct values of $z^{1/n}$. The n roots of a complex quantity are uniformly spaced around a circle, with radius $r^{1/n}$, in the complex plane in a symmetric fashion.

Example Find the three cube roots of -8. Here $r = 8$, $\theta = \pi$. The roots are $z_0 = 2(\cos\pi/3 + i\sin\pi/3) = 1 + i\sqrt{3}$, $z_1 = 2(\cos\pi + i\sin\pi) = -2$, $z_2 = 2(\cos 5\pi/3 + i\sin 5\pi/3) = 1 - i\sqrt{3}$.

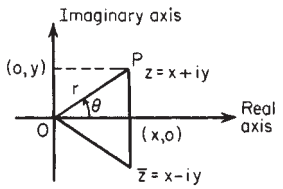

FIG. 3-46 Complex plane.

ELEMENTARY COMPLEX FUNCTIONS

Polynomials A polynomial in z, $a_nz^n + a_{n-1}z^{n-1} + \cdots + a_0$, where n is a positive integer, is simply a sum of complex numbers times integral powers of z which have already been defined. Every polynomial of degree n has precisely n complex roots provided each multiple root of multiplicity m is counted m times.

Exponential Functions The exponential function e^z is defined by the equation $e^z = e^{x+iy} = e^x \cdot e^{iy} = e^z(\cos y + i\sin y)$. Properties: $e^0 = 1$; $e^{z_1} \cdot e^{z_2} = e^{z_1+z_2}$; $e^{z_1}/e^{z_2} = e^{z_1-z_2}$; $e^{z+2k\pi i} = e^z$.

Trigonometric Functions $\sin z = (e^{iz} - e^{-iz})/2i$; $\cos z = (e^{iz} + e^{-iz})/2$; $\tan z = \sin z/\cos z$; $\cot z = \cos z/\sin z$; $\sec z = 1/\cos z$; $\csc z = 1/\sin z$. Fundamental identities for these functions are the same as their real counterparts. Thus $\cos^2 z + \sin^2 z = 1$, $\cos(z_1 \pm z_2) = \cos z_1\cos z_2 \mp \sin z_1\sin z_2$, $\sin(z_1 \pm z_2) = \sin z_1\cos z_2 \pm \cos z_1\sin z_2$. The sine and cosine of z are periodic functions of period 2π; thus $\sin(z + 2\pi) = \sin z$. For computation purposes $\sin z = \sin(x + iy) = \sin x\cosh y + i\cos x\sinh y$, where $\sin x$, $\cosh y$, etc., are the real trigonometric and hyperbolic functions. Similarly, $\cos z = \cos x\cosh y - i\sin x\sinh y$. If $x = 0$ in the results given, $\cos iy = \cosh y$, $\sin iy = i\sinh y$.

Example Find all solutions of $\sin z = 3$. From previous data $\sin z = \sin x\cosh y + i\cos x\sinh y = 3$. Equating real and imaginary parts $\sin x\cosh y = 3$, $\cos x\sinh y = 0$. The second equation can hold for $y = 0$ or for $x = \pi/2, 3\pi/2, \ldots$. If $y = 0$, $\cosh 0 = 1$ and $\sin x = 3$ is impossible for real x. Therefore, $x = \pm\pi/2, \pm3\pi/2, \ldots \pm(2n + 1)\pi/2$, $n = 0, \pm1, \pm2, \ldots$. However, $\sin 3\pi/2 = -1$ and $\cosh y \ge 1$. Hence $x = \pi/2, 5\pi/2, \ldots$. The solution is $z = [(4n + 1)\pi]/2 + i\cosh^{-1}3$, $n = 0, 1, 2, 3, \ldots$.

Example Find all solutions of $e^z = -i$. $e^z = e^x(\cos y + i\sin y) = -i$. Equating real and imaginary parts gives $e^x\cos y = 0$, $e^x\sin y = -1$. From the first $y = \pm\pi/2, \pm3\pi/2, \ldots$. But $e^x > 0$. Therefore, $y = 3\pi/2, 7\pi/2, -\pi/2, \ldots$. Then $x = 0$. The solution is $z = i[(4n + 3)\pi]/2$.

Two important facets of these functions should be recognized. First, the $\sin z$ is *unbounded;* and, second, e^z takes *all* complex values *except* 0.

Hyperbolic Functions $\sinh z = (e^z - e^{-z})/2$; $\cosh z = (e^z + e^{-z})/2$; $\tanh z = \sinh z/\cosh z$; $\coth z = \cosh z/\sinh z$; $\operatorname{csch} z = 1/\sinh z$; $\operatorname{sech} z = 1/\cosh z$. Identities are: $\cosh^2 z - \sinh^2 z = 1$; $\sinh(z_1 + z_2) = \sinh z_1\cosh z_2 + \cosh z_1\sinh z_2$; $\cosh(z_1 + z_2) = \cosh z_1\cosh z_2 + \sinh z_1\sinh z_2$; $\cosh z + \sinh z = e^z$; $\cosh z - \sinh z = e^{-z}$. The hyperbolic sine and hyperbolic cosine are periodic functions with the imaginary period $2\pi i$. That is, $\sinh(z + 2\pi i) = \sinh z$.

Logarithms The logarithm of z, $\log z = \log|z| + i(\theta + 2n\pi)$, where $\log|z|$ is taken to the base e and θ is the principal argument of z, that is, the particular argument lying in the interval $0 \le \theta < 2\pi$. The logarithm of z is infinitely many valued. If $n = 0$, the resulting logarithm is called the principal value. The familiar laws $\log z_1z_2 = \log z_1 + \log z_2$, $\log z_1/z_2 = \log z_1 - \log z_2$, $\log z^n = n\log z$ hold for the principal value.

Example $\log(1+i) = \log\sqrt{2} + i\left(\frac{\pi}{4} + 2n\pi\right)$.

General powers of z are defined by $z^\alpha = e^{\alpha \log z}$. Since $\log z$ is infinitely many valued, so too is z^α unless α is a rational number.

DeMoivre's formula can be derived from properties of e^z.

$$z^n = r^n(\cos\theta + i\sin\theta)^n = r^n(\cos n\theta + i\sin n\theta)$$

Thus $$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta$$

Example $i^i = e^{i\log i} = e^{i[\log|i| + i(\pi/2 + 2n\pi)]} = e^{-(\pi/2 + 2n\pi)}$. Thus i^i is real with principal value $(n = 0) = e^{-\pi/2}$.

Example $(\sqrt{2})^{1+i} = e^{(1+i)\log\sqrt{2}} = e^{\log\sqrt{2}}$. $e^{i\log\sqrt{2}} = \sqrt{2}$. $(\cos\log\sqrt{2} + i\sin\log\sqrt{2}) = \sqrt{2}[\cos(0.3466) + i\sin(0.3466)]$.

Inverse Trigonmetric Functions $\cos^{-1} z = -i\log(z \pm \sqrt{z^2-1})$; $\sin^{-1} z = -i\log(iz \pm \sqrt{1-z^2})$; $\tan^{-1} z = \frac{i}{2}\log\left(\frac{i+z}{i-z}\right)$. These functions are infinitely many valued.

Inverse Hyperbolic Functions $\cosh^{-1} z = \log(z \pm \sqrt{z^2-1})$; $\sinh^{-1} z = \log(z \pm \sqrt{z^2+1})$; $\tanh^{-1} z = \frac{1}{2}\log\left(\frac{1+z}{1-z}\right)$.

COMPLEX FUNCTIONS (ANALYTIC)

In the real-number system a greater than $b(a > b)$ and b less than $c(b < c)$ define an order relation. These relations have no meaning for complex numbers. The absolute value is used for ordering. Some important relations follow: $|z| \geq x$; $|z| \geq y$; $|z_1 \pm z_2| \leq |z_1| + |z_2|$; $|z_1 - z_2| \geq ||z_1| - |z_2||$; $|z| \geq (|x| + |y|)/\sqrt{2}$. Parts of the complex plane, commonly called **regions** or **domains,** are described by using inequalities.

Example $|z-3| \leq 5$. This is equivalent to $\sqrt{(x-3)^2 + y^2} \leq 5$, which is the set of all points within and on the circle, centered at $x = 3$, $y = 0$ of radius 5.

Example $|z-1| \leq x$ represents the set of all points inside and on the parabola $2x = y^2 + 1$ or, equivalently, $2x \geq y^2 + 1$.

Functions of a Complex Variable If $z = x + iy$, $w = u + iv$ and if for each value of z in some region of the complex plane one or more values of w are defined, then w is said to be a function of z, $w = f(z)$. Some of these functions have already been discussed, e.g., $\sin z$, $\log z$. All functions are reducible to the form $w = u(x, y) + iv(x, y)$, where u, v are real functions of the real variables x and y.

Example $z^3 = (x+iy)^3 = x^3 + 3x^2(iy) + 3x(iy)^2 + (iy)^3 = (x^3 - 3xy^2) + i(3x^2y - y^3)$.

Example $\cos z = \cos x\cosh y - i\sin x\sinh y$.

Differentiation The *derivative* of $w = f(z)$ is

$$\frac{dw}{dz} = \lim_{\Delta z \to 0} \frac{f(z+\Delta z) - f(z)}{\Delta z}$$

and for the derivative to exist the limit must be the same no matter how Δz approaches zero. If w_1, w_2 are differentiable functions of z, the following rules apply:

$$\frac{d(w_1 \pm w_2)}{dz} = \frac{dw_1}{dz} \pm \frac{dw_2}{dz} \qquad \frac{d(w_1 w_2)}{dz} = w_2\frac{dw_1}{dz} + w_1\frac{dw_2}{dz}$$

$$\frac{d(w_1/w_2)}{dz} = \frac{w_2(dw_1/dz) - w_1(dw_2/dz)}{w_2^2}$$

and $$\frac{dw_1^n}{dz} = nw_1^{n-1}\frac{dw_1}{dz}$$

For $w = f(z)$ to be differentiable, it is necessary that $\partial u/\partial x = \partial v/\partial y$ and $\partial v/\partial x = -\partial u/\partial y$. The last two equations are called the Cauchy-Riemann equations. The derivative

$$\frac{dw}{dz} = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} - i\frac{\partial u}{\partial y}$$

If $f(z)$ possesses a derivative at z_o and at every point in some neighborhood of z_0, then $f(z)$ is said to be analytic at z_0. If the Cauchy-Riemann equations are satisfied and

$$u,\ v,\ \frac{\partial u}{\partial x},\ \frac{\partial u}{\partial y},\ \frac{\partial v}{\partial x},\ \frac{\partial v}{\partial y}$$

are continuous in a region of the complex plane, then $f(z)$ is analytic in that region.

Example $w = z\bar{z} = x^2 + y^2$. Here $u = x^2 + y^2$, $v = 0$. $\partial u/\partial x = 2x$, $\partial u/\partial y = 2y$, $\partial v/\partial x = \partial v/\partial y = 0$. These are continuous everywhere, but the Cauchy-Riemann equations hold only at the origin. Therefore, w is nowhere analytic, but it is differentiable at $z = 0$ only.

Example $w = e^{\bar{z}} = e^x\cos y + ie^x\sin y$. $u = e^x\cos y$, $v = e^x\sin y$. $\partial u/\partial x = e^x\cos y$, $\partial u/\partial y = -e^x\sin y$, $\partial v/\partial x = e^x\sin y$, $\partial v/\partial y = e^x\cos y$. The continuity and Cauchy-Riemann requirements are satisfied for all finite z. Hence $e^{\bar{z}}$ is analytic (except at ∞) and $dw/\partial z = \partial u/\partial x + i(\partial v/\partial x) = e^{\bar{z}}$.

Example $w = \frac{1}{z} = \frac{x - iy}{x^2 + y^2} = \frac{x}{x^2+y^2} - i\frac{y}{x^2+y^2}$

It is easy to see that dw/dz exists except at $z = 0$. Thus $1/z$ is analytic except at $z = 0$.

Singular Points If $f(z)$ is analytic in a region except at certain points, those points are called singular points.

Example $1/z$ has a singular point at zero.

Example $\tan z$ has singular points at $z = \pm(2n+1)(\pi/2)$, $n = 0, 1, 2, \ldots$.

The derivatives of the common functions, given earlier, are the same as their real counterparts.

Example $(d/dz)(\log z) = 1/z$, $(d/dz)(\sin z) = \cos z$.

Harmonic Functions Both the *real* and the *imaginary* parts of any analytic function $f = u + iv$ satisfy Laplace's equation $\partial^2\phi/\partial x^2 + \partial^2\phi/\partial y^2 = 0$. A function which possesses continuous second partial derivatives and satisfies Laplace's equation is called a harmonic function.

Example $e^{\bar{z}} = e^x\cos y + ie^x\sin y$. $u = e^x\cos y$, $\partial u/\partial x = e^x\cos y$, $\partial^2 u/\partial x^2 = e^x\cos y$, $\partial u/\partial y = -e^x\sin y$, $\partial^2 u/\partial y^2 = -e^x\cos y$. Clearly $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = 0$. Similarly, $v = e^x\sin y$ is also harmonic.

If $w = u + iv$ is analytic, the curves $u(x, y) = c$ and $v(x, y) = k$ intersect at right angles, if $w'(z) \neq 0$.

Example $z^3 = (x^3 - 3xy^2) + i(3x^2y - y^3)$. Set $u = x^3 - 3xy^2 = c$, $v = 3x^2y - y^3 = k$. By implicit differentiation there results, respectively, $dy/dx = (x^2 - y^2)/2xy$, $dy/dx = 2xy/(y^2 - x^2)$, which are clearly negative reciprocals, the condition for perpendicularity.

Integration In much of the work with complex variables a simple extension of integration called line or curvilinear integration is of fundamental importance. Since any complex line integral can be expressed in terms of real line integrals, we define only real line integrals. Let $F(x,y)$ be a real, continuous function of x and y and c be any continuous curve of finite length joining the points A and B (Fig. 3-47). $F(x,y)$ is not related to the curve c. Divide c up into n segments, Δs_i, whose projection on the x axis is Δx_i and on the y axis is Δy_i. Let (ε_i, η_i) be the coordinates of an arbitrary point on Δs_i. The limits of the sums

$$\lim_{\Delta s_i \to 0} \sum_{i=1}^{n} F(\varepsilon_i, \eta_i)\,\Delta s_i = \int_c F(x, y)\,ds$$

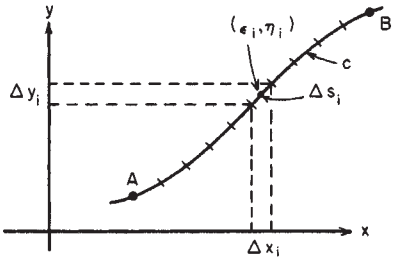

FIG. 3-47 Line integral.

$$\lim_{\Delta s_i \to 0} \sum_{i=1}^{n} F(\varepsilon_i, \eta_i)\,\Delta x_i = \int_c F(x, y)\,dx$$

$$\lim_{\Delta s_i \to 0} \sum_{i=1}^{n} F(\varepsilon_i, \eta_i)\,\Delta y_i = \int_c F(x, y)\,dy$$

are known as line integrals. Much of the initial strangeness of these integrals will vanish if it be observed that the ordinary definite integral $\int_a^b f(x)\,dx$ is just a line integral in which the curve c is a line segment on the x axis and $F(x, y)$ is a function of x alone. The evaluation of line integrals can be reduced to evaluation of ordinary integrals.

Example $\int_c y(1+x)\,dy$, where c: $y = 1 - x^2$ from $(-1, 0)$ to $(1, 0)$. Clearly $y = 1 - x^2$, $dy = -2x\,dx$. Thus $\int_c y(1+x)\,dy = -2\int_{-1}^{1}(1-x^2)(1+x)x\,dx = -8/15$.

Example $\int_c x^2 y\,ds$, c is the square whose vertices are $(0, 0)$, $(1, 0)$, $(1, 1)$, $(0, 1)$. $ds = \sqrt{dx^2 + dy^2}$. When $dx = 0$, $ds = dy$. From $(0, 0)$ to $(1, 0)$, $y = 0$, $dy = 0$. Similar arguments for the other sides give

$$\int_c x^2 y\,ds = \int_0^1 0\,x^2\,dx + \int_0^1 y\,dy + \int_1^0 x^2\,dx + \int_1^0 0\,y\,dy = \tfrac{1}{2} - \tfrac{1}{3} = \tfrac{1}{6}$$

Let $f(z)$ be any function of z, analytic or not, and c any curve as above. The complex integral is calculated as $\int_c f(z)\,dz = \int_c (u\,dx - v\,dy) + i\int_c (v\,dx + u\,dy)$, where $f(z) = u(x, y) + iv(x, y)$. Properties of line integrals are the same as those for ordinary integrals. That is, $\int_c [f(z) \pm g(z)]\,dz = \int_c f(z)\,dz \pm \int_c g(z)\,dz$; $\int_c kf(z)\,dz = k\int_c f(z)\,dz$ for any constant k, etc.

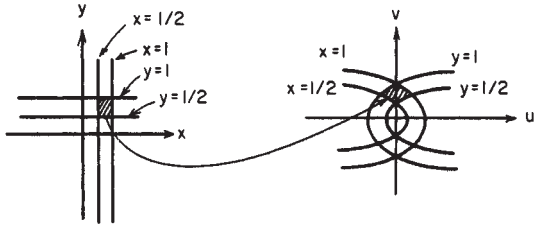

FIG. 3-48 Conformal transformation.

Example $\int_c (x^2 + iy)\,dz$ along c: $y = x$, 0 to $1 + i$. This becomes

$$\int_c (x^2 + iy)\,dz = \int_c (x^2\,dx - y\,dy)$$

$$+ i\int_c (y\,dx + x^2\,dy) = \int_0^1 x^2\,dx - \int_0^1 x\,dx + i\int_0^1 x\,dx + i\int_0^1 x^2\,dx = -\tfrac{1}{6} + 5i/6$$

Conformal Mapping Every function of a complex variable $w = f(z) = u(x, y) + iv(x, y)$ transforms the x, y plane into the u, v plane in some manner. A conformal transformation is one in which angles between curves are preserved in *magnitude* and *sense*. Every analytic function, except at those points where $f'(z) = 0$, is a conformal transformation. See Fig. 3-48.

Example $w = z^2$. $u + iv = (x^2 - y^2) + 2ixy$ or $u = x^2 - y^2$, $v = 2xy$. These are the transformation equations between the (x, y) and (u, v) planes. Lines parallel to the x axis, $y = c_1$ map into curves in the u, v plane with parametric equations $u = x^2 - c_1^2$, $v = 2c_1x$. Eliminating x, $u = (v^2/4c_1^2) - c_1^2$, which represents a family of parabolas with the origin of the w plane as focus, the line $v = 0$ as axis and opening to the right. Similar arguments apply to $x = c_2$.

The principles of complex variables are useful in the solution of a variety of applied problems. See the references for additional information.

DIFFERENTIAL EQUATIONS

REFERENCES: *Ordinary Differential Equations:* Elementary level, 41, 44, 62, 81, 204, 236, 263. Intermediate level, 30, 43, 144. Theory and Advanced topics, 252. Applications, 9, 263. Partial Differential Equations: Elementary level and solution methods, 9, 41, 61, 72, 144, 156, 229. Theory and advanced level, 79, 220, 240.

See also "Numerical Analysis and Approximate Methods" and "General References: References for General and Specific Topics—Advanced Engineering Mathematics" for additional references on topics in ordinary and partial differential equations.

The natural laws in any scientific or technological field are not regarded as precise and definitive until they have been expressed in mathematical form. Such a form, often an equation, is a relation between the quantity of interest, say, product yield, and independent variables such as time and temperature upon which yield depends. When it happens that this equation involves, besides the function itself, one or more of its derivatives it is called a differential equation.

Example The homogeneous bimolecular reaction $A + B \xrightarrow{k} C$ is characterized by the differential equation $dx/dt = k(a - x)(b - x)$, where a = initial concentration of A, b = initial concentration of B, and $x = x(t)$ = concentration of C as a function of time t.

Example The differential equation of heat conduction in a moving fluid with velocity components v_x, v_y is

$$\frac{\partial u}{\partial t} + v_x\frac{\partial u}{\partial x} + v_y\frac{\partial u}{\partial y} = \frac{K}{\rho c_p}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

where $u = u(x, y, t)$ = temperature, K = thermal conductivity, ρ = density, and c_p = specific heat at constant pressure.

ORDINARY DIFFERENTIAL EQUATIONS

When the function involved in the equation depends upon only one variable, its derivatives are ordinary derivatives and the differential equation is called an ordinary differential equation. When the function depends upon several independent variables, then the equation is called a partial differential equation. The theories of ordinary and partial differential equations are quite different. In almost every respect the latter is more difficult.

Whichever the type, a differential equation is said to be of nth order if it involves derivatives of order n but no higher. The equation in the first example is of first order and that in the second example of second order. The degree of a differential equation is the power to which the derivative of the highest order is raised after the equation has been cleared of fractions and radicals in the dependent variable and its derivatives.

A relation between the variables, involving no derivatives, is called a solution of the differential equation if this relation, when substituted in the equation, satisfies the equation. A solution of an ordinary differential equation which includes the maximum possible number of "arbitrary" constants is called the **general solution.** The maximum number of "arbitrary" constants is exactly equal to the order of the dif-

ferential equation. If any set of specific values of the constants is chosen, the result is called a **particular solution.**

Example The general solution of $(d^2x/dt^2) + k^2x = 0$ is $x = A \cos kt + B \sin kt$, where A, B are arbitrary constants. A particular solution is $x = \frac{1}{2} \cos kt + 3 \sin kt$.

In the case of some equations still other solutions exist called singular solutions. A **singular solution** *is any solution of the differential equation which is not included in the general solution.*

Example $y = x(dy/dx) - \frac{1}{4}(dy/dx)^2$ has the general solution $y = cx - \frac{1}{4}c^2$, where c is an arbitrary constant; $y = x^2$ is a singular solution, as is easily verified.

ORDINARY DIFFERENTIAL EQUATIONS OF THE FIRST ORDER

Equations with Separable Variables Every differential equation of the first order and of the first degree can be written in the form $M(x, y)\,dx + N(x, y)\,dy = 0$. If the equation can be transformed so that M does not involve y and N does not involve x, then the variables are said to be separated. The solution can then be obtained by **quadrature,** which means that $y = \int f(x)\,dx + c$, which may or may not be expressible in simpler form.

Example Two liquids A and B are boiling together in a vessel. Experimentally it is found that the ratio of the rates at which A and B are evaporating at any time is proportional to the ratio of the amount of A (say, x) to the amount of B (say, y) still in the liquid state. This physical law is expressible as $(dy/dt)/(dx/dt) = ky/x$ or $dy/dx = ky/x$, where k is a proportionality constant. This equation may be written $dy/y = k(dx/x)$, in which the variables are separated. The solution is $\ln y = k \ln x + \ln c$ or $y = cx^k$.

Exact Equations The equation $M(x, y)\,dx + N(x, y)\,dy = 0$ is exact if and only if $\partial M/\partial y = \partial N/\partial x$. In this case there exists a function $w = f(x, y)$ such that $\partial f/\partial x = M$, $\partial f/\partial y = N$, and $f(x, y) = C$ is the required solution. $f(x, y)$ is found as follows: treat y as though it were constant and evaluate $\int M(x, y)\,dx$. Then treat x as though it were constant and evaluate $\int N(x, y)\,dy$. The sum of all unlike terms in these two integrals (including no repetitions) is $f(x, y)$.

Example $(2xy - \cos x)\,dx + (x^2 - 1)\,dy = 0$ is exact for $\partial M/\partial y = 2x$, $\partial N/\partial x = 2x$. $\int M\,dx = \int (2xy - \cos x)\,dx = x^2y - \sin x$, $\int N\,dy = \int (x^2 - 1)\,dy = x^2y - y$. The solution is $x^2y - \sin x - y = C$, as may easily be verified.

Linear Equations A differential equation is said to be linear when it is of first degree in the dependent variable and its derivatives. The general linear first-order differential equation has the form $dy/dx + P(x)y = Q(x)$. Its general solution is

$$y = e^{-\int P\,dx}\left[\int Qe^{\int P\,dx}\,dx + C\right]$$

Example A tank initially holds 200 gal of a salt solution in which 100 lb is dissolved. Six gallons of brine containing 4 lb of salt run into the tank per minute. If mixing is perfect and the output rate is 4 gal/min, what is the amount A of salt in the tank at time t? The differential equation of A is $dA/dt + [1/(100 + t)]A = 4$. Its general solution is $A = 2(100 + t) + C/(100 + t)$. At $t = 0$, $A = 100$; so the particular solution is $A = 2(100 + t) - 10^4/(100 + t)$.

ORDINARY DIFFERENTIAL EQUATIONS OF HIGHER ORDER

The higher-order differential equations, especially those of order 2, are of great importance because of physical situations describable by them.

Equation $y^{(n)} = f(x)$ Such a differential equation can be solved by n integrations. The solution will contain n arbitrary constants.

Linear Differential Equations with Constant Coefficients and Right-Hand Member Zero (Homogeneous) The solution of $y'' + ay' + by = 0$ depends upon the nature of the roots of the characteristic equation $m^2 + am + b = 0$ obtained by substituting the trial solution $y = e^{mx}$ in the equation.

Distinct Real Roots If the roots of the characteristic equation are distinct real roots, r_1 and r_2, say, the solution is $y = Ae^{r_1x} + Be^{r_2x}$, where A and B are arbitrary constants.

Example $y'' + 4y' + 3 = 0$. The characteristic equation is $m^2 + 4m + 3 = 0$. The roots are -3 and -1, and the general solution is $y = Ae^{-3x} + Be^{-x}$.

Multiple Real Roots If $r_1 = r_2$, the solution of the differential equation is $y = e^{r_1x}(A + Bx)$.

Example $y'' + 4y + 4 = 0$. The characteristic equation is $m^2 + 4m + 4 = 0$ with roots -2 and -2. The solution is $y = e^{-2x}(A + Bx)$.

Complex Roots If the characteristic roots are $p \pm iq$, then the solution is $y = e^{px}(A \cos qx + B \sin qx)$.

Example The differential equation $My'' + Ay' + ky = 0$ represents the vibration of a linear system of mass M, spring constant k, and damping constant A. If $A < 2\sqrt{kM}$, the roots of the characteristic equation

$$Mm^2 + Am + k = 0 \text{ are complex } -\frac{A}{2M} \pm i\sqrt{\frac{k}{M} - \left(\frac{A}{2M}\right)^2}$$

and the solution is $y = e^{-(At/2M)}$

$$\left\{c_1 \cos\left(\sqrt{\frac{k}{M} - \left(\frac{A}{2M}\right)^2}\right)t + c_2 \sin\left(\sqrt{\frac{k}{M} - \left(\frac{A}{2M}\right)^2}\right)t\right\}$$

This solution is oscillatory, representing undercritical damping.

All these results generalize to homogeneous linear differential equations with constant coefficients of order higher than 2. These equations (especially of order 2) have been much used because of the ease of solution. Oscillations, electric circuits, diffusion processes, and heat-flow problems are a few examples for which such equations are useful.

Second-Order Equations: Dependent Variable Missing Such an equation is of the form

$$F\left(x, \frac{dy}{dx}, \frac{d^2y}{dx^2}\right) = 0$$

It can be reduced to a first-order equation by substituting $p = dy/dx$ and $dp/dx = d^2y/dx^2$.

Second-Order Equations: Independent Variable Missing Such an equation is of the form

$$F\left(y, \frac{dy}{dx}, \frac{d^2y}{dx^2}\right) = 0$$

Set $$\frac{dy}{dx} = p, \quad \frac{d^2y}{dx^2} = p\,\frac{dp}{dy}$$

The result is a first-order equation in p,

$$F\left(y, p, p\,\frac{dp}{dy}\right) = 0$$

Example The capillary curve for one vertical plate is given by

$$\frac{d^2y}{dx^2} = \frac{4y}{c^2}\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{3/2}$$

Its solution by this technique is

$$x + \sqrt{c^2 - y^2} - \sqrt{c^2 - h_0^2} = \frac{c}{2}\left(\cosh^{-1}\frac{c}{y} - \cosh^{-1}\frac{c}{h_0}\right)$$

where c, h_0 are physical constants.

Example The equation governing chemical reaction in a porous catalyst in plane geometry of thickness L is

$$D\,\frac{d^2c}{dx^2} = k\,f(c), \quad \frac{dc}{dx}(0) = 0, \quad c(L) = c_0$$

where D is a diffusion coefficient, k is a reaction rate parameter, c is the concentration, $k\,f(c)$ is the rate of reaction, and c_0 is the concentration at the boundary. Making the substitution gives

$$p\,\frac{dp}{dc} = \frac{k}{D}f(c)$$

Integrating gives $$\frac{p^2}{2} = \frac{k}{D}\int_{c(0)}^{c} f(c)\,dc$$

If the reaction is very fast, $c(0) \approx 0$ and the average reaction rate is related to $p(L)$. See Ref. 106. This variable is given by

$$p(L) = \left[\frac{2k}{D}\int_0^{c_0} f(c)\,dc\right]^{1/2}$$

Thus, the average reaction rate can be calculated without solving the complete problem.

Linear Nonhomogeneous Differential Equations

Linear Differential Equations Right-Hand Member f(x) ≠ 0 Again the specific remarks for $y'' + ay' + by = f(x)$ apply to differential equations of similar type but higher order. We shall discuss two general methods.

Method of Undetermined Coefficients Use of this method is limited to equations exhibiting both constant coefficients and particular forms of the function $f(x)$. In most cases $f(x)$ will be a sum or product of functions of the type constant, x^n (n a positive integer), e^{mx}, $\cos kx$, $\sin kx$. When this is the case, the solution of the equation is $y = H(x) + P(x)$, where $H(x)$ is a solution of the homogeneous equations found by the method of the preceding subsection and $P(x)$ is a particular integral found by using the following table subject to these conditions: (1) When $f(x)$ consists of the sum of several terms, the appropriate form of $P(x)$ is the sum of the particular integrals corresponding to these terms individually. (2) When a term in any of the trial integrals listed is already a part of the homogeneous solution, the indicated form of the particular integral is multiplied by x.

Form of Particular Integral

If $f(x)$ is	Then $P(x)$ is
a (constant)	A (constant)
ax^n	$A_n x^n + A_{n-1}x^{n-1} + \cdots A_1 x + A_0$
ae^{rx}	Be^{rx}
$c\cos kx$ $d\sin kx$	$A\cos kx + B\sin kx$
$gx^n e^{rx}\cos kx$ $hx^n e^{rx}\sin kx$	$(A_n x^n + \cdots + A_0)e^{rx}\cos kx + (B_n x^n + \cdots + B_0)e^{rx}\sin kx$

Since the form of the particular integral is known, the constants may be evaluated by substitution in the differential equation.

Example $y'' + 2y' + y = 3e^{2x} - \cos x + x^3$. The characteristic equation is $(m+1)^2 = 0$ so that the homogeneous solution is $y = (c_1 + c_2 x)e^{-x}$. To find a particular solution we use the trial solution from the table, $y = a_1 e^{2x} + a_2\cos x + a_3\sin x + a_4 x^3 + a_5 x^2 + a_6 x + a_7$. Substituting this in the differential equation collecting and equating like terms, there results $a_1 = \frac{1}{3}$, $a_2 = 0$, $a_3 = -\frac{1}{2}$, $a_4 = 1$, $a_5 = -6$, $a_6 = 18$, and $a_7 = -24$. The solution is $y = (c_1 + c_2 x)e^{-x} + \frac{1}{3}e^{2x} - \frac{1}{2}\sin x + x^3 - 6x^2 + 18x - 24$.

Method of Variation of Parameters This method is applicable to any linear equation. The technique is developed for a second-order equation but immediately extends to higher order. Let the equation be $y'' + a(x)y' + b(x)y = R(x)$ and let the solution of the homogeneous equation, found by some method, be $y = c_1 f_1(x) + c_2 f_2(x)$. It is now assumed that a particular integral of the differential equation is of the form $P(x) = uf_1 + vf_2$ where u, v are functions of x to be determined by two equations. One equation results from the requirement that $uf_1 + vf_2$ satisfy the differential equation, and the other is a degree of freedom open to the analyst. The best choice proves to be

$$u'f_1 + v'f_2 = 0 \qquad \text{and} \qquad u'f_1' + v'f_2' = R(x)$$

Then $$u' = \frac{du}{dx} = -\frac{f_2}{f_1 f_2' - f_2 f_1'}R(x)$$

$$v' = \frac{dv}{dx} = \frac{f_1}{f_1 f_2' - f_2 f_1'}R(x)$$

and since f_1, f_2, and R are known u, v may be found by direct integration.

Example $(1-x^2)\dfrac{d^2y}{dx^2} - \dfrac{1}{x}\dfrac{dy}{dx} = x$. The homogeneous equation

$$(1-x^2)\frac{d^2y}{dx^2} - \frac{1}{x}\frac{dy}{dx} = 0$$

reduces to $$\frac{dp}{p} = \frac{dx}{x(1-x^2)}$$

when we set $dy/dx = p$. Upon integrating twice, $y = c_1\sqrt{x^2-1} + c_2$ is the homogeneous solution. Now assume that the particular solution has the form $y = u\sqrt{x^2-1} + v$. The equations for u and v become

$$u' = du/dx = \sqrt{x^2-1}$$

$$v' = \frac{dv}{dx} = 1 - x^2$$

so that

$$u = \frac{1}{2}[x\sqrt{x^2-1} - \ln(x + \sqrt{x^2-1})] \qquad \text{and} \qquad v = x - x^3/3.$$

The complete solution is

$$y = c_1\sqrt{x^2-1} + c_2 + \frac{x}{2} - \frac{x^3}{6} - \frac{1}{2}\sqrt{x^2-1}\ln(x + \sqrt{x^2-1}).$$

Perturbation Methods If the ordinary differential equation has a parameter that is small and is not multiplying the highest derivative, perturbation methods can give solutions for small values of the parameter.

Example Consider the differential equation for reaction and diffusion in a catalyst; the reaction is second order: $c'' = ac^2$, $c'(0) = 0$, $c(1) = 1$. The solution is expanded in the following Taylor series in a.

$$c(x, a) = c_0(x) + ac_1(x) + a^2 c_2(x) + \cdots$$

The goal is to find equations governing the functions $\{c_i(x)\}$ and solve them. Substitution into the equations gives the following equations:

$$c_0''(x) + a\,c_1''(x) + a^2 c_2''(x) + \cdots = a[c_0(x) + ac_1(x) + a^2 c_2(x) + \cdots]^2$$

$$c_0'(0) + ac_1'(0) + a^2 c_2'(0) + \cdots = 0$$

$$c_0(1) + ac_1(1) + a^2 c_2(1) + \cdots = 1$$

Like terms in powers of a are collected to form the individual problems.

$$c_0'' = 0, \quad c_0'(0) = 0, \quad c_0(1) = 1$$

$$c_1'' = c_0^2, \quad c_1'(0) = 0, \quad c_1(1) = 0$$

$$c_2'' = 2c_0 c_1, \quad c_2'(0) = 0, \quad c_2(1) = 0$$

The solution proceeds in turn.

$$c_0(x) = 1, \quad c_1(x) = \frac{(x^2-1)}{2}, \quad c_2(x) = \frac{5 - 6x^2 + x^4}{12}$$

SPECIAL DIFFERENTIAL EQUATIONS (SEE REF. 1)

Euler's Equation The linear equation $x^n y^{(n)} + a_1 x^{n-1}y^{(n-1)} + \cdots + a_{n-1}xy' + a_n y = R(x)$ can be reduced to a linear equation with constant coefficients by the change of variable $x = e^t$. To solve the homogeneous equation substitute $y = x^r$ into it, cancel the powers of x, which are the same for all terms, and solve the resulting polynomial for r. In case of multiple or complex roots there results the form $y = x^r(\log x)^r$ and $y = x^\alpha[\cos(\beta\log x) + i\sin(\beta\log x)]$.

Example Solve $x^2y'' - 2y = 0$. By setting $y = x^r$, $x^r[r(r-1) - 2] = 0$. The roots of $r^2 - r - 2 = 0$ are $r = 2, -1$. The general solution is $y = Ax^2 + B/x$.

The equation $(ax+b)^n y^{(n)} + a_1(ax+b)^{n-1}y^{(n-1)} + \cdots + a_n y = R(x)$ can be reduced to the Euler form by the substitution $ax + b = z$. It may be treated without change of variable, the homogeneous equation having solutions of the form $y = (ax+b)^r$.

Bessel's Equation The linear equation $x^2(d^2y/dx^2) + (1 - 2\alpha)x(dy/dx) + [\beta^2\gamma^2 x^{2\gamma} + (\alpha^2 - p^2\gamma^2)]y = 0$ is the general Bessel equation. By series methods, not to be discussed here, this equation can be shown to have the solution

$$y = Ax^\alpha J_p(\beta x^\gamma) + Bx^\alpha J_{-p}(\beta x^\gamma) \qquad p \text{ not an integer or zero}$$

$$y = Ax^\alpha J_p(\beta x^\gamma) + Bx^\alpha Y_p(\beta x^\gamma) \qquad p \text{ an integer}$$

where $J_p(x) = \left(\frac{x}{2}\right)^p \sum_{k=0}^{\infty} \frac{(-1)^k (x/2)^{2k}}{k!\Gamma(p+k+1)}$

$$J_{-p}(x) = \left(\frac{x}{2}\right)^{-p} \sum_{k=0}^{\infty} \frac{(-1)^k (x/2)^{2k}}{k!\Gamma(k+1-p)} \quad p \text{ not an integer}$$

$$\Gamma(n) = \int_0^{\infty} x^{n-1} e^{-x}\, dx \quad n > 0$$

is the gamma function. For p an integer

$$J_p(x) = \left(\frac{x}{2}\right)^p \sum_{k=0}^{\infty} \frac{(-1)^k (x/2)^{2k}}{k!(p+k)!}$$

(Bessel function of the first kind of order p)

$$Y_p(x) = \frac{[J_p(x)\cos(p\pi) - J_{-p}(x)]}{\sin(p\pi)}$$

(replace right-hand side by limiting value if P is an integer or zero).

The series converge for all x. Much of the importance of Bessel's equation and Bessel functions lies in the fact that the solutions of numerous linear differential equations can be expressed in terms of them.

Example $d^2y/dx^2 + [9x - (63/4x^2)]y = 0$. In general form this is $x^2(d^2y/dx^2) + (9x^3 - 6\frac{3}{4})y = 0$. Thus $\alpha = \frac{1}{2}$, $\gamma = \frac{3}{2}$; $\beta = 2$, $p = \frac{3}{3}$. The solution is (since $p \neq$ integer) $y = Ax^{1/2}J_{3/3}(2x^{3/2}) + Bx^{1/2}J_{-3/3}(2x^{3/2})$. Tables are available for the evaluation of many of these functions.

Example The heat flow through a wedge-shaped fin is characterized by the equation $x^2(d^2y/dx^2) + x(dy/dx) - a^2xy = 0$, where $y = T - T_{air}$, α is a combination of physical constants, and x = distance from fin end. By comparing this with the standard equation, there results $\alpha = 0$, $p = 0$, $\gamma = \frac{1}{2}$, $\beta^2 = -4a^2$ or $\beta = 2ai$. The solution is $y = AJ_0(2ai\sqrt{x}) + BY_0(2ai\sqrt{x})$.

Legendre's Equation The Legendre equation $(1 - x^2)y'' - 2xy' + n(n+1)y = 0$, $n \geq 0$, has the solution $y = Au_n(x) + Bv_n(x)$ for n not an integer where

$$u_n(x) = 1 - \frac{n(n+1)}{2!}x^2 + \frac{n(n-2)(n+1)(n+3)}{4!}x^4 - \frac{n(n-2)(n-4)(n+1)(n+3)(n+5)}{6!}x^6 + \cdots$$

$$v_n(x) = x - \frac{(n-1)(n+2)}{3!}x^3 + \frac{(n-1)(n-3)(n+2)(n+4)}{5!}x^5 - \cdots$$

If n is an even integer or zero, u_n is a polynomial in x. If n is an odd integer, then v_n is a polynomial. The interval of convergence for the series is $-1 < x < 1$. If n is an integer, set

$$P_n(x) = \frac{u_n(x)}{u_n(1)} \ (n \text{ even or zero}), \quad P_n = \frac{v_n(x)}{v_n(1)} \ (n \text{ odd})$$

The polynomials P_n are the so-called Legendre polynomials, $P_0(x) = 1$, $P_1(x) = x$, $P_2(x) = \frac{1}{2}(3x^2 - 1)$, $P_3(x) = \frac{1}{2}(5x^3 - 3x)$, ...

Laguerre's Equation The Laguerre equation $x(d^2y/dx^2) + (c - x)(dy/dx) - ay = 0$ is satisfied by the confluent hypergeometric function. See Refs. 1 and 173.

Hermite's Equation The Hermite equation $y'' - 2xy' + 2ny = 0$ is satisfied by the Hermite polynomial of degree n, $y = AH_n(x)$ if n is a positive integer or zero. $H_0(x) = 1$, $H_1(x) = 2x$, $H_2(x) = 4x^2 - 2$, $H_3(x) = 8x^3 - 12x$, $H_4(x) = 16x^4 - 48x^2 + 12$, $H_{r+1}(x) = 2xH_r(x) - 2rH_{r-1}(x)$.

Example $y'' - 2xy' + 6y = 0$. Here $n = 3$; so $y = AH_3 = A(8x^3 - 12x)$ is a solution.

Chebyshev's Equation The equation $(1 - x^2)y'' - xy' + n^2y = 0$ for n a positive integer or zero is satisfied by the nth Chebyshev polynomial $y = AT_n(x)$. $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3x$, $T_4(x) = 8x^4 - 8x^2 + 1$; $T_{r+1}(x) = 2xT_r(x) - T_{r-1}(x)$.

Example $(1 - x^2)y'' - xy' + 36y = 0$. Here $n = 6$. A solution is $y = T_6(x) = 2xT_5(x) - T_4(x) = 2x(2xT_4 - T_3) - T_4 = 32x^6 - 48x^4 + 18x^2 - 1$. Further details on these special equations and others can be found in the literature.

PARTIAL DIFFERENTIAL EQUATIONS

The analysis of situations involving two or more independent variables frequently results in a partial differential equation.

Example The equation $\partial T/\partial t = K(\partial^2 T/\partial x^2)$ represents the unsteady one-dimensional conduction of heat.

Example The equation for the unsteady transverse motion of a uniform beam clamped at the ends is

$$\frac{\partial^4 y}{\partial x^4} + \frac{\rho}{EI}\frac{\partial^2 y}{\partial t^2} = 0$$

Example The expansion of a gas behind a piston is characterized by the simultaneous equations

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{c^2}{\rho}\frac{\partial \rho}{\partial x} = 0 \quad \text{and} \quad \frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \rho\frac{\partial u}{\partial x} = 0$$

Example The heating of a diathermanous solid is characterized by the equation $\alpha(\partial^2\theta/\partial x^2) + \beta e^{-x_c} = \partial\theta/\partial t$.

The partial differential equation $\partial^2 f/\partial x\,\partial y = 0$ can be solved by two integrations yielding the solution $f = g(x) + h(y)$, where $g(x)$ and $h(y)$ are arbitrary differentiable functions. This result is an example of the fact that the general solution of partial differential equations involves arbitrary functions in contrast to the solution of ordinary differential equations, which involve only arbitrary constants. A number of methods are available for finding the general solution of a partial differential equation. In most applications of partial differential equations the general solution is of limited use. In such applications the solution of a partial differential equation must satisfy both the equation and certain auxiliary conditions called **initial** and/or **boundary** conditions, which are dictated by the problem. Examples of these include those in which the wall temperature is a fixed constant $T(x_0) = T_0$, there is no diffusion across a nonpermeable wall, and the like. In ordinary differential equations these auxiliary conditions allow definite numbers to be assigned to the constants of integration. In partial differential equations the boundary conditions demand that the arbitrary functions resulting from integration assume specific forms. Except for a few cases (some first-order equations, D'Alembert's solution of the wave equation, and others) a procedure which first determines the arbitrary functions and then specializes them to fit the boundary conditions is usually not feasible. A more fruitful attack is to determine directly a set of particular solutions and then combine them so that the boundary conditions are satisfied. The only area in which much analysis has been accomplished is for linear homogeneous partial differential equations. Such equations have the property that if $f_1, f_2, \ldots, f_n, \ldots$ are individually solutions, then the function $f = \sum_{i=1}^{\infty} f_i$ is also a solution, provided the series converges and is differentiable up to the order (termwise) of the equation.

Partial Differential Equations of Second and Higher Order Many of the applications to scientific problems fall naturally into partial differential equations of second order, although there are important exceptions in elasticity, vibration theory, and elsewhere.

A second-order differential equation can be written as

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = f$$

where a, b, c, and f depend upon x, y, u, $\partial u/\partial x$, and $\partial u/\partial y$. This equation is hyperbolic, parabolic, or elliptic, depending on whether the discriminant $b^2 - 4ac$ is >0, =0, or <0, respectively. Since a, b, c, and f depend on the solution, the type of equation can be different at different x and y locations. If the equation is hyperbolic, discontinuities can be propagated. See Refs. 11, 79, 105, 159, and 192.

Phenomena of **propagation** such as vibrations are characterized by equations of "hyperbolic" type which are essentially different in their properties from other classes such as those which describe equilibrium (elliptic) or unsteady diffusion and heat transfer (parabolic). Prototypes are as follows:

Elliptic Laplace's equation $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = 0$ and Poisson's equation $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = g(x, y)$ do not contain the variable time

explicitly and consequently represent equilibrium configurations. Laplace's equation is satisfied by static electric or magnetic potential at points free from electric charges or magnetic poles. Other important functions satisfying Laplace's equation are the velocity potential of the irrotational motion of an incompressible fluid, used in hydrodynamics; the steady temperature at points in a homogeneous solid, and the steady state of diffusion through a homogeneous body. The gravitational potential V at points occupied by mass of density d satisfies Poisson's equation $\partial^2V/\partial x^2 + \partial^2V/\partial y^2 + \partial^2V/\partial z^2 = -4\pi d$.

Parabolic The heat equation $\partial T/\partial t = \partial^2T/\partial x^2 + \partial^2T/\partial y^2$ represents nonequilibrium or unsteady states of heat conduction and diffusion.

Hyperbolic The wave equation $\partial^2u/\partial t^2 = c^2(\partial^2u/\partial x^2 + \partial^2u/\partial y^2)$ represents wave propagation of many varied types.

Quasilinear first-order differential equations are like

$$a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} = f$$

where a, b, and f depend on x, y, and u, with $a^2 + b^2 \neq 0$. This equation can be solved using the method of characteristics, which writes the solution in terms of a parameter s, which defines a path for the characteristic.

$$\frac{dx}{ds} = a, \quad \frac{dy}{ds} = b, \quad \frac{du}{ds} = f$$

These equations are integrated from some initial conditions. For a specified value of s, the value of x and y shows the location where the solution is u. The equation is semilinear if a and b depend just on x and y (and not u), and the equation is linear if a, b, and f all depend on x and y, but not u. Such equations give rise to shock propagation, and conditions have been derived to deduce the presence of shocks, Ref. 245. For further information, see Refs. 79, 159, 192, and 245.

An example of a linear hyperbolic equation is the advection equation for flow of contaminants when the x and y velocity components are u and v, respectively.

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} + v\frac{\partial c}{\partial y} = 0$$

The equations for flow and adsorption in a packed bed or chromatography column give a quasilinear equation.

$$\phi\frac{\partial c}{\partial t} + \phi\, u\frac{\partial c}{\partial x} + (1-\phi)\frac{df}{dc}\frac{\partial c}{\partial t} = 0$$

Here $n = f(c)$ is the relation between concentration on the adsorbent and fluid concentration.

The solution of problems involving partial differential equations often revolves about an attempt to reduce the partial differential equation to one or more ordinary differential equations. The solutions of the ordinary differential equations are then combined (if possible) so that the boundary conditions as well as the original partial differential equation are simultaneously satisfied. Three of these techniques are illustrated.

Similarity Variables The physical meaning of the term "similarity" relates to internal similitude, or self-similitude. Thus, similar solutions in boundary-layer flow over a horizontal flat plate are those for which the horizontal component of velocity u has the property that two velocity profiles located at different coordinates x differ only by a scale factor. The mathematical interpretation of the term similarity is a transformation of variables carried out so that a reduction in the number of independent variables is achieved. There are essentially two methods for finding similarity variables, "separation of variables" (not the classical concept) and the use of "continuous transformation groups." The basic theory is available in Ames (see the references).

Example The equation $\partial\theta/\partial x = (A/y)(\partial^2\theta/\partial y^2)$ with the boundary conditions $\theta = 0$ at $x = 0$, $y > 0$; $\theta = 0$ at $y = \infty$, $x > 0$; $\theta = 1$ at $y = 0$, $x > 0$ represents the nondimensional temperature θ of a fluid moving past an infinitely wide flat plate immersed in the fluid. Turbulent transfer is neglected, as is molecular transport except in the y direction. It is now assumed that the equation and the boundary conditions can be satisfied by a solution of the form $\theta = f(y/x^n) = f(u)$, where $\theta = 0$ at $u = \infty$ and $\theta = 1$ at $u = 0$. The purpose here is to replace the independent variables x and y by the single variable u when it is hoped that a value of n exists which will allow x and y to be completely eliminated in the equation. In this case since $u = y/x^n$, there results after some calculation $\partial\theta/\partial x = -(nu/x)(d\theta/du)$, $\partial^2\theta/\partial y^2 = (1/x^{2n})(d^2\theta/du^2)$, and when these are substituted in the equation, $-(1/x)nu(d\theta/du) = (1/x^{3n})(A/u)(d^2\theta/du^2)$. For this to be a function of u only, choose $n = \frac{1}{3}$. There results $(d^2\theta/du^2) + (u^3/3A)(d\theta/du) = 0$. Two integrations and use of the boundary conditions for this ordinary differential equation give the solution

$$\theta = \int_u^\infty \exp(-u^3/9A)\, du \Big/ \int_0^\infty \exp(-u^3/9A)\, du$$

Group Method The type of transformation can be deduced using group theory. For a complete exposition, see Refs. 9, 12, and 145; a shortened version is in Ref. 106. Basically, a similarity transformation should be considered when one of the independent variables has no physical scale (perhaps it goes to infinity). The boundary conditions must also simplify (and combine) since each transformation leads to a differential equation with one fewer independent variable.

Example A similarity variable is found for the problem

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\left(D(c)\frac{\partial c}{\partial x}\right), \quad c(0,t) = 1, \quad c(\infty,t) = 0, \quad c(x,0) = 0$$

Note that the length dimension goes to infinity, so that there is no length scale in the problem statement; this is a clue to try a similarity transformation. The transformation examined here is

$$\bar{t} = a^\alpha t, \quad \bar{x} = a^\beta x, \quad \bar{c} = a^\gamma c$$

With this substitution, the equation becomes

$$a^{\alpha-\gamma}\frac{\partial \bar{c}}{\partial \bar{t}} = a^{2\beta-\gamma}\frac{\partial}{\partial \bar{x}}\left[D(a^{-\gamma}\bar{c})\frac{\partial \bar{c}}{\partial \bar{x}}\right]$$

Group theory says a system is conformally invariant if it has the same form in the new variables; here, that is

$$\gamma = 0, \quad \alpha - \gamma = 2\beta - \gamma, \quad \text{or } \alpha = 2\beta$$

The invariants are

$$\eta = \frac{x}{t^\delta}, \quad \delta = \frac{\beta}{\alpha}$$

and the solution is

$$c(x,t) = f(\eta)t^{\gamma/\alpha}$$

We can take $\gamma = 0$ and $\delta = \beta/\alpha = \frac{1}{2}$. Note that the boundary conditions combine because the point $x = \infty$ and $t = 0$ give the same value of η and the conditions on c at $x = \infty$ and $t = 0$ are the same. We thus make the transformation

$$\eta = \frac{x}{\sqrt{4D_0 t}}, \quad c(x,t) = f(\eta)$$

The use of the 4 and D_0 makes the analysis below simpler. The result is

$$\frac{d}{d\eta}\left[D(c)\frac{df}{d\eta}\right] + 2\eta\frac{df}{d\eta} = 0, \quad f(0) = 1, \quad f(\infty) = 0$$

Thus, we solve a two-point boundary value problem instead of a partial differential equation. When the diffusivity is constant, the solution is the error function, a tabulated function.

$$c(x,t) = 1 - \operatorname{erf}\eta = \operatorname{erfc}\eta$$

$$\operatorname{erf}\eta = \int_0^\eta e^{-\xi^2}\, d\xi \Big/ \int_0^\infty e^{-\xi^2}\, d\xi$$

Separation of Variables This is a powerful, well-utilized method which is applicable in certain circumstances. It consists of assuming that the solution for a partial differential equation has the form $U = f(x)g(y)$. If it is then possible to obtain an ordinary differential equation on one side of the equation depending only on x and on the other side only on y, the partial differential equation is said to be separable in the variables x, y. If this is the case, one side of the equation is a function of x alone and the other of y alone. The two can be equal only if each is a constant, say λ. Thus the problem has again been reduced to the solution of ordinary differential equations.

Example Laplace's equation $\partial^2V/\partial x^2 + \partial^2V/\partial y^2 = 0$ plus the boundary conditions $V(0, y) = 0$, $V(l, y) = 0$, $V(x, \infty) = 0$, $V(x, 0) = f(x)$ represents the steady-state potential in a thin plate (in z direction) of infinite extent in the y direction

and of width l in the x direction. A potential $f(x)$ is impressed (at $y = 0$) from $x = 0$ to $x = 1$, and the sides are grounded. To obtain a solution of this boundary-value problem assume $V(x, y) = f(x)g(y)$. Substitution in the differential equation yields $f''(x)g(y) + f(x)g''(y) = 0$, or $g''(y)/g(y) = -f''(x)/f(x) = \lambda^2$ (say). This system becomes $g''(y) - \lambda^2 g(y) = 0$ and $f''(x) + \lambda^2 f(x) = 0$. The solutions of these ordinary differential equations are respectively $g(y) = Ae^{\lambda y} + Be^{-\lambda y}$, $f(x) = C \sin \lambda x + D \cos \lambda x$. Then $f(x)g(y) = (Ae^{\lambda y} + Be^{-\lambda y})\ (C \sin \lambda x + D \cos \lambda x)$. Now $V(0, y) = 0$ so that $f(0)g(y) = (Ae^{\lambda y} + Be^{-\lambda y})\ D \equiv 0$ for all y. Hence $D = 0$. The solution then has the form $\sin \lambda x\ (Ae^{\lambda y} + Be^{-\lambda y})$ where the multiplicative constant C has been eliminated. Since $V(l, y) = 0$, $\sin \lambda l(Ae^{\lambda y} + Be^{-\lambda y}) \equiv 0$. Clearly the bracketed function of y is not zero, for the solution would then be the identically zero solution. Hence $\sin \lambda l = 0$ or $\lambda_n = n\pi/l$, $n = 1, 2, \ldots$ where $\lambda_n = n$th eigenvalue.

The solution now has the form $\sin (n\pi x/l)(Ae^{n\pi y/l} + Be^{-n\pi y/l})$. Since $V(x, \infty) = 0$, A must be taken to be zero because e^y becomes arbitrarily large as $y \to \infty$. The solution then reads $B_n \sin (n\pi x/l)e^{-n\pi y/l}$, where B_n is the multiplicative constant. The differential equation is linear and homogeneous so that $\sum_{n=1}^{\infty} B_n e^{-n\pi y/l} \sin (n\pi x/l)$ is also a solution. Satisfaction of the last boundary condition is ensured by taking

$$B_n = \frac{2}{l}\int_0^l f(x) \sin (n\pi x/l)\, dx = \text{Fourier sine coefficients of } f(x)$$

Further, convergence and differentiability of this series are established quite easily. Thus the solution is

$$V(x, y) = \sum_{n=1}^{\infty} B_n e^{-n\pi y/l} \sin \frac{n\pi x}{l}$$

Example The diffusion problem

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\left(D(c)\,\frac{\partial c}{\partial x}\right), \quad c(0, t) = 1, \quad c(\infty, t) = 0, \quad c(x, 0) = 0$$

can be solved by separation of variables. First transform the problem so that the boundary conditions are homogeneous (having zeroes on the right-hand side). Let

$$c(x, t) = 1 - x + u(x, t)$$

Then $u(x, t)$ satisfies

$$\frac{\partial u}{\partial t} = D\,\frac{\partial^2 u}{\partial x^2}, \quad u(x, 0) = x - 1, \quad u(0, t) = 0, \quad u(1, t) = 0$$

Assume a solution of the form $u(x, t) = X(x)\ T(t)$, which gives

$$X\,\frac{dT}{dt} = D\,T\,\frac{d^2X}{dx^2}$$

Since both sides are constant, this gives the following ordinary differential equations to solve.

$$\frac{1}{D\,T}\,\frac{dT}{dt} = -\lambda, \quad \frac{1}{X}\,\frac{d^2X}{dx^2} = -\lambda$$

The solution of these is

$$T = A\,e^{-\lambda Dt}, \quad X = B \cos \sqrt{\lambda}\,x + E \sin \sqrt{\lambda}\,x$$

The combined solution for $u(x,t)$ is

$$u = A\,(B \cos \sqrt{\lambda}\,x + E \sin \sqrt{\lambda}\,x)\,e^{-\lambda Dt}$$

Apply the boundary condition that $u(0,t) = 0$ to give $B = 0$. Then the solution is

$$u = A\,(\sin \sqrt{\lambda}\,x)e^{-\lambda Dt}$$

where the multiplicative constant E has been eliminated. Apply the boundary condition at $x = L$.

$$0 = A\,(\sin \sqrt{\lambda}\,L)e^{-\lambda Dt}$$

This can be satisfied by choosing $A = 0$, which gives no solution. However, it can also be satisfied by choosing λ such that

$$\sin \sqrt{\lambda}\,L = 0, \quad \sqrt{\lambda}\,L = n\,\pi$$

Thus

$$\lambda = \frac{n^2\pi^2}{L^2}$$

The combined solution can now be written as

$$u = A\left(\frac{\sin n\pi x}{L}\right)e^{-n^2\pi^2 Dt/L^2}$$

Since the initial condition must be satisfied, we use an infinite series of these functions.

$$u = \sum_{n=1}^{\infty} A_n\left(\frac{\sin n\pi x}{L}\right)e^{-n^2\pi^2 Dt/L^2}$$

At $t = 0$, we satisfy the initial condition.

$$x - 1 = \sum_{n=1}^{\infty} A_n\left(\frac{\sin n\pi x}{L}\right)$$

This is done by multiplying the equation by

$$\frac{\sin m\pi x}{L}$$

and integrating over x: $0 \to L$. (This is the same as minimizing the mean-square error of the initial condition.) This gives

$$\frac{A_m L}{2} = \int_0^L (x - 1) \sin m\pi x\, dx$$

which completes the solution.

Integral-Transform Method A number of integral transforms are used in the solution of differential equations. Only one, the Laplace transform, will be discussed here [for others, see "Integral Transforms (Operational Methods)"]. The one-sided Laplace transform indicated by $L[f(t)]$ is defined by the equation $L[f(t)] = \int_0^{\infty} f(t)e^{-st}\, dt$. It has numerous important properties. The ones of interest here are $L[f'(t)] = sL[f(t)] - f(0)$; $L[f''(t)] = s^2L[f(t)] - sf(0) - f'(0)$; $L[f^{(n)}(t)] = s^nL[f(t)] - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - f^{(n-1)}(0)$ for ordinary derivatives. For partial derivatives an indication of which variable is being transformed avoids confusion. Thus, if

$$y = y(x, t), \quad L_t\left[\frac{\partial y}{\partial t}\right] = sL[y(x, t)] - y(x, 0)$$

whereas

$$L_t\left[\frac{\partial y}{\partial x}\right] = \frac{dL_t[y(x, t)]}{dx}$$

since $L[y(x, t)]$ is "really" only a function of x. Otherwise the results are similar. These facts coupled with the linearity of the transform, i.e., $L[af(t) + bg(t)] = aL[f(t)] + bL[g(t)]$, make it a useful device in solving some linear differential equations. Its use reduces the solution of ordinary differential equations to the solution of algebraic equations for $L[y]$. The solution of partial differential equations is reduced to the solution of ordinary differential equations. In both situations the inverse transform must be obtained either from tables, of which there are several, or by use of complex inversion methods.

Example The equation $\partial c/\partial t = D(\partial^2 c/\partial x^2)$ represents the diffusion in a semi-infinite medium, $x \geq 0$. Under the boundary conditions $c(0, t) = c_0$, $c(x, 0) = 0$ find a solution of the diffusion equation. By taking the Laplace transform of both sides with respect to t,

$$\int_0^{\infty} e^{-st}\,\frac{\partial^2 c}{\partial x^2}\,dt = \frac{1}{D}\int_0^{\infty} e^{-st}\,\frac{\partial c}{\partial t}\,dt$$

or

$$\frac{d^2F}{dx^2} = (1/D)sF - c(x, 0) = \frac{sF}{D}$$

where $F(x, s) = L_t[c(x, t)]$. Hence

$$\frac{d^2F}{dx^2} - \left(\frac{s}{D}\right)F = 0$$

The other boundary condition transforms into $F(0, s) = c_0/s$. Finally the solution of the ordinary differential equation for F subject to $F(0, s) = c_0/s$ and F remains finite as $x \to \infty$ is $F(x, s) = (c_0/s)e^{-\sqrt{s/D}x}$. Reference to a table shows that the function having this as its Laplace transform is

$$c(x, t) = c_0\left[1 - \frac{2}{\sqrt{\pi}}\int_0^{x/2\sqrt{Dt}} e^{-u^2}\,du\right]$$

Matched-Asymptotic Expansions Sometimes the coefficient in front of the highest derivative is a small number. Special perturbation techniques can then be used, provided the proper scaling laws are found. See Refs. 32, 170, and 180.

DIFFERENCE EQUATIONS

REFERENCES: 30, 43.

Certain situations are such that the independent variable does not vary continuously but has meaning only for discrete values. Typical illustrations occur in the stagewise processes found in chemical engineering such as distillation, staged extraction systems, and absorption columns. In each of these the operation is characterized by a finite between-stage change of the dependent variable in which the independent variable is the integral number of the stage. The importance of difference equations is twofold: (1) to analyze problems of the type described and (2) to obtain approximate solutions of problems which lead, in their formulation, to differential equations. In this subsection only problems of analysis are considered; the application to approximate solutions is considered under "Numerical Analysis and Approximate Methods."

ELEMENTS OF THE CALCULUS OF FINITE DIFFERENCES

Let $y = f(x)$ be defined for discrete equidistant values of x, which will be denoted by x_n. The corresponding value of y will be written $y_n = f(x_n)$. The first forward difference of $f(x)$ denoted by $\Delta f(x) = f(x + h) - f(x)$ where $h = x_n - x_{n-1}$ = interval length.

Example Let $f(x) = x^2$. Then $\Delta f(x) = (x + h)^2 - x^2 = 2hx + h^2$.

The second forward difference is obtained by taking the difference of the first; thus $\Delta\Delta f(x) = \Delta^2 f(x) = \Delta f(x + h) - \Delta f(x) = f(x + 2h) - 2f(x + h) + f(x)$.

Example $f(x) = x^2$, $\Delta^2 f(x) = \Delta[\Delta f(x)] = \Delta 2hx + \Delta h^2 = 2h(x + h) - 2hx + h^2 - h^2 = 2h^2$.

Similarly the nth forward difference is defined by the relation $\Delta^n f(x) = \Delta[\Delta^{n-1} f(x)]$. Other difference relations are also quite useful. Some of these are $\nabla f(x) = f(x) - f(x - h)$, which is called the backward difference, and $\delta f(x) = f[x + (h/2)] - f[x - (h/2)]$, called the central difference. Some properties of the operator Δ are quite important. If C is any constant, $\Delta C = 0$; if $f(x)$ is any function of period h, $\Delta f(x) = 0$ (in fact, periodic functions of period h play the same role here as constants do in the differential calculus); $\Delta[f(x) + g(x)] = \Delta f(x) + \Delta g(x)$; $\Delta^m[\Delta^n f(x)] = \Delta^{m+n} f(x)$; $\Delta[f(x)g(x)] = f(x)\,\Delta g(x) + g(x + h)\,\Delta f(x)$

$$\Delta\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)\,\Delta f(x) - f(x)\,\Delta g(x)}{g(x)g(x + h)}$$

Example $\Delta(x \sin x) = x\Delta \sin x + \sin(x + h)\,\Delta x = 2x \sin(h/2) \cos[x + (h/2)] + h \sin(x + h)$.

DIFFERENCE EQUATIONS

A difference equation is a relation between the differences and the independent variable, $\phi(\Delta^n y, \Delta^{n-1} y, \ldots, \Delta y, y, x) = 0$, where ϕ is some given function. The general case in which the interval between the successive points is any real number h, instead of 1, can be reduced to that with interval size 1 by the substitution $x = hx'$. Hence all further difference-equation work will assume the interval size between successive points is 1.

Example $f(x + 1) - (\alpha + 1)f(x) + \alpha f(x - 1) = 0$. Common notation usually is $y_x = f(x)$. This equation is then written $y_{x+1} - (\alpha + 1)y_x + \alpha y_{x-1} = 0$.

Example $y_{x+2} + 2y_x y_{x+1} + y_x = x^2$.

Example $y_{x+1} - y_x = 2^x$.

The order of the difference equation is the difference between the largest and smallest arguments when written in the form of the second example. The first and second examples are both of order 2, while the third example is of order 1. A linear difference equation involves no products or other nonlinear functions of the dependent variable and its differences. The first and third examples are linear, while the second example is nonlinear.

A solution of a difference equation is a relation between the variables which satisfies the equation. If the difference equation is of order n, the general solution involves n arbitrary constants. The techniques for solving difference equations resemble techniques used for differential equations.

Equation $\Delta^n y = a$ The solution of $\Delta^n y = a$, where a is a constant, is a polynomial of degree n plus an arbitrary periodic function of period 1. That is, $y = (ax^n/n!) + c_1 x^{n-1} + c_2 x^{n-2} + \cdots + c_n + f(x)$, where $f(x + 1) = f(x)$.

Example $\Delta^3 y = 6$. The solution is $y = x^3 + c_1 x^2 + c_2 x + c_3 + f(x)$; c_1, c_2, c_3 are arbitrary constants, and $f(x)$ is an arbitrary periodic function of period 1.

Equation $y_{x+1} - y_x = \phi(x)$ This equation states that the first difference of the unknown function is equal to the given function $\phi(x)$. The solution by analogy with solving the differential equation $dy/dx = \phi(x)$ by integration is obtained by "finite integration" or summation. When there are only a finite number of data points, this is easily accomplished by writing $y_x = y_0 + \sum_{t=1}^{x} \phi(t - 1)$, where the data points are numbered from 1 to x. This is the only situation considered here.

Examples If $\phi(x) = 1$, $y_x = x$. If $\phi(x) = x$, $y_x = [x(x - 1)]/2$. If $\phi(x) = a^x$, $a \neq 0$, $y_x = a^x/(a - 1)$. In all cases $y_0 = 0$.

Other examples may be evaluated by using summation, that is, $y_2 = y_1 + \phi(1)$, $y_3 = y_2 + \phi(2) = y_1 + \phi(1) + \phi(2)$, $y_4 = y_3 + \phi(3) = y_1 + \phi(1) + \phi(2) + \phi(3)$, $\ldots$, $y_x = y_1 + \sum_{t=1}^{x-1} \phi(t)$.

Example $y_{x+1} - ry_x = 1$, r constant, $x > 0$ and $y_0 = 1$. $y_1 = 1 + r$, $y_2 = 1 + r + r^2$, $\ldots$, $y_x = 1 + r + \cdots + r^x = (1 - r^{x+1})/(1 - r)$ for $r \neq 1$ and $y_x = 1 + x$ for $r = 1$.

Linear Difference Equations The linear difference equation of order n has the form $P_n y_{x+n} + P_{n-1} y_{x+n-1} + \cdots + P_1 y_{x+1} + P_0 y_x = Q(x)$ with $P_n \neq 0$ and $P_0 \neq 0$ and P_j; $j = 0, \ldots, n$ are functions of x.

Constant Coefficient and Q(x) = 0 (Homogeneous) The solution is obtained by trying a solution of the form $y_x = c\beta^x$. When this trial solution is substituted in the difference equation, a polynomial of degree n results for β. If the solutions of this polynomial are denoted by $\beta_1, \beta_2, \ldots, \beta_n$ then the following cases result: (1) if all the β_j's are real and unequal, the solution is $y_x = \sum_{j=1}^{n} c_j \beta_j^x$, where the $c_1, \ldots, c_n$ are arbitrary constants; (2) if the roots are real and repeated, say, β_j has multiplicity m, then the partial solution corresponding to β_j is $\beta_j^x(c_1 + c_2 x + \cdots + c_m x^{m-1})$; (3) if the roots are complex conjugates, say, $a + ib = pe^{i\theta}$ and $a - ib = pe^{-i\theta}$, the partial solution corresponding to this pair is $p^x(c_1 \cos \theta x + c_2 \sin \theta x)$; and (4) if the roots are multiple complex conjugates, say, $a + ib = pe^{i\theta}$ and $a - ib = pe^{-i\theta}$ are m-fold, then the partial solution corresponding to these is $p^x[(c_1 + c_2 x + \cdots + c_m x^{m-1}) \cos \theta x + (d_1 + d_2 x + \cdots + d_m x^{m-1}) \sin \theta x]$.

Example The equation $y_{x+1} - (\alpha + 1)y_x + \alpha y_{x-1} = 0$, $y_0 = c_0$ and $y_{m+1} = x_{m+1}/k$ represents the steady-state composition of transferable material in the raffinate stream of a staged countercurrent liquid-liquid extraction system. Clearly y is a function of the stage number x. α is a combination of system constants. By using the trial solution $y_x = c\beta^x$, there results $\beta^2 - (\alpha + 1)\beta + \alpha = 0$, so that $\beta_1 = 1$, $\beta_2 = \alpha$. The general solution is $y_x = c_1 + c_2\alpha^x$. By using the side conditions, $c_1 = c_0 - c_2$, $c_2 = (y_{m+1} - c_0)/(\alpha^{m+1} - 1)$. The desired solution is $(y_x - c_0)/(y_{m+1} - c_0) = (\alpha^x - 1)/(\alpha^{m+1} - 1)$.

Example $y_{x+3} - 3y_{x+2} + 4y_x = 0$. By setting $y_x = c\beta^x$, there results $\beta^3 - 3\beta^2 + 4 = 0$ or $\beta_1 = -1$, $\beta_2 = 2$, $\beta_3 = 2$. The general solution is $y_x = c_1(-1)^x + 2^x(c_2 + c_3 x)$.

Example $y_{x+1} - 2y_x + 2y_{x-1} = 0$. $\beta_1 = 1 + i$, $\beta_2 = 1 - i$. $p = \sqrt{1 + 1} = \sqrt{2}$, $\theta = \pi/4$. The solution is $y_x = 2^{x/2}[c_1 \cos(x\pi/4) + c_2 \sin(x\pi/4)]$.

Constant Coefficients and Q(x) ≠ 0 (Nonhomogeneous) In this case the general solution is found by first obtaining the homoge-

neous solution, say, y_x^H and adding to it any particular solution with $Q(x) \neq 0$, say, y_x^p. There are several means of obtaining the particular solution.

Method of Undetermined Coefficients If $Q(x)$ is a product or linear combination of products of the functions e^{bx}, a^x, x^p (p a positive integer or zero) $\cos cx$ and $\sin cx$, this method may be used. The "families" $[a^x]$, $[e^{bx}]$, $[\sin cx, \cos cx]$ and $[x^p, x^{p-1}, \ldots, x, 1]$ are defined for each of the above functions in the following way: The family of a term f_x is the set of all functions of which f_x and all operations of the form a^{x+y}, $\cos c(x+y)$, $\sin c(x+y)$, $(x+y)^p$ on f_x and their linear combinations result in. The technique involves the following steps: (1) Solve the homogeneous system. (2) Construct the family of each term. (3) If the family has no representative in the homogeneous solution, assume y_x^p is a linear combination of the families of each term and determine the constants so that the equation is satisfied. (4) If a family has a representative in the homogeneous solution, multiply each member of the family by the smallest integral power of x for which all such representatives are removed and revert to step 3.

Example $y_{x+1} - 3y_x + 2y_{x-1} = 1 + a^x$. $a \neq 0$. The homogeneous solution is $y_x^H = c_1 + c_2 2^x$. The family of 1 is 1 and of a^x is a^x. However, 1 is a solution of the homogeneous system. Therefore, try $y_x^p = Ax + Ba^x$. Substituting in the equation there results

$$y_x = c_1 + c_2 2^x - x + \frac{a}{(a-1)(a-2)} a^x a \neq 1, a \neq 2$$

If $a = 1$, $y_x = c_1 + c_2 2^x - 2x$. If $a = 2$, $y_x = c_1 + c_2 2^x - x + x2^x$.

Example The family of $x^2 3^x$ is $[x^2 3^x, x3^x, 3^x]$.

Method of Variation of Parameters This technique is applicable to general linear difference equations. It is illustrated for the second-order system $y_{x+2} + Ay_{x+1} + By_x = \phi(x)$. Assume that the homogeneous solution has been found by some technique and write $y_x^H = c_1 u_x + c_2 v_x$. Assume that a particular solution $y_x^p = D_x u_x + E_x v_x$. E_x and D_x can be found by solving the equations:

$$E_{x+1} - E_x = \frac{u_{x+1}\phi(x)}{u_{x+1}v_{x+2} - u_{x+2}v_{x+1}}$$

$$D_{x+1} - D_x = \frac{v_{x+1}\phi(x)}{v_{x+1}u_{x+2} - v_{x+2}u_{x+1}}$$

by summation. The general solution is then $y_x = y_x^p + y_x^H$.

Variable Coefficients The method of variation of parameters applies equally well to the linear difference equation with variable coefficients. Techniques are therefore needed to solve the homogeneous system with variable coefficients.

Equation $y_{x+1} - a_x y_x = 0$ By assuming that this equation is valid for $x \geq 0$ and $y_0 = c$, the solution is $y_x = c \prod_{n=1}^{x} a_{n-1}$.

Example $y_{x+1} + \frac{x+2}{x+1} y_x = 0$. The solution is

$$y_x = c \prod_{n=1}^{x} \left(-\frac{n+1}{n}\right) = c(-1)^x \frac{2}{1} \cdot \frac{3}{2} \cdots \frac{x+1}{x} = (-1)^x c(x+1)$$

Example $y_{x+1} - xy_x = 0$. The solution is $y_x = c(x-1)!$

Reduction of Order If one homogeneous solution, say, u_x, can be found by inspection or otherwise, an equation of lower order can be obtained by the substitution $v_x = y_x/u_x$. The resultant equation must be satisfied by v_x = constant or $\Delta v_x = 0$. Thus the equation will be of reduced order if the new variable $U_x = \Delta(y_x/u_x)$ is introduced.

Example $(x+2)y_{x+2} - (x+3)y_{x+1} + y_x = 0$. By observation $u_x = 1$ is a solution. Set $U_x = \Delta y_x = y_{x+1} - y_x$. There results $(x+2)U_{x+1} - U_x = 0$, which is of degree one lower than the original equation. The complete solution for y_x is finally

$$y_x = c_0 \sum_{n=0}^{x} \frac{1}{n!} + c_1$$

Factorization If the difference equation can be factored, then the general solution can be obtained by solving two or more successive equations of lower order. Consider $y_{x+2} + A_x y_{x+1} + B_x y_x = \phi(x)$. If there exists a_x, b_x such that $a_x + b_x = -A_x$ and $a_x b_x = B_x$, then the difference equation may be written $y_{x+2} - (a_x + b_x)\, y_{x+1} + a_x b_x y_x = \phi(x)$. First solve $U_{x+1} - b_x U_x = \phi(x)$ and then $y_{x+1} - a_x y_x = U_x$.

Example $y_{x+2} - (2x+1)y_{x+1} + (x^2 + x)y_x = 0$. Set $a_x = x$, $b_x = x + 1$. Solve $u_{x+1} - (x+1)u_x = 0$ and then $y_{x+1} - xy_x = u_x$.

Substitution If it is possible to rearrange a difference equation so that it takes the form $af_{x+2}y_{x+2} + bf_{x+1}y_{x+1} + cf_x y_x = \phi(x)$ with a, b, c constants, then the substitution $u_x = f_x y_x$ reduces the equation to one with constant coefficients.

Example $(x+2)^2 y_{x+2} - 3(x+1)^2 y_{x+1} + 2x^2 y_x = 0$. Set $u_x = x^2 y_x$. The equation becomes $u_{x+2} - 3u_{x+1} + 2u_x = 0$, which is linear and easily solved by previous methods.

The substitution $u_x = y_x/f_x$ reduces $af_x f_{x+1} y_{x+z} + bf_x f_{x+2} y_{x+1} + cf_{x+1} f_{x+2} y_x = \phi(x)$ to an equation with constant coefficients.

Example $x(x+1)y_{x+2} + 3x(x+2)y_{x+1} - 4(x+1)(x+2)y_x = x$. Set $u_x = y_x/f_x = y_x/x$. Then $y_x = xu_x$, $y_{x+1} = (x+1)u_{x+1}$ and $y_{x+2} = (x+2)u_{x+2}$. Substitution in the equation yields $x(x+1)(x+2)u_{x+2} + 3x(x+2)(x+1)u_{x+1} - 4x(x+1)(x+2)u_x = x$ or $u_{x+2} + 3u_{x+1} - 4u_x = 1/(x+1)(x+2)$, which is a linear equation with constant coefficients.

Nonlinear Difference Equations: Riccati Difference Equation The Riccati equation $y_{x+1}y_x + ay_{x+1} + by_x + c = 0$ is a nonlinear difference equation which can be solved by reduction to linear form. Set $y = z + h$. The equation becomes $z_{x+1}z_x + (h+a)z_{x+1} + (h+b)z_x + h^2 + (a+b)h + c = 0$. If h is selected as a root of $h^2 + (a+b)h + c = 0$ and the equation is divided by $z_{x+1}z_x$ there results $[(h+b)/z_{x+1}] + [(h+a)/z_x] + 1 = 0$. This is a linear equation with constant coefficients. The solution is

$$y_x = h + \frac{1}{c\left[-\dfrac{a+h}{b+h}\right]^x - \dfrac{1}{(a+h)+(b+h)}}$$

Example This equation is obtained in distillation problems, among others, in which the number of theoretical plates is required. If the relative volatility is assumed to be constant, the plates are theoretically perfect, and the molal liquid and vapor rates are constant, then a material balance around the nth plate of the enriching section yields a Riccati difference equation.

INTEGRAL EQUATIONS

REFERENCES: 75, 79, 105, 195, 273. See also "Numerical Analysis and Approximate Methods."

An integral equation is any equation in which the unknown function appears under the sign of integration and possibly outside the sign of integration. If derivatives of the dependent variable appear elsewhere in the equation, the equation is said to be integrodifferential.

CLASSIFICATION OF INTEGRAL EQUATIONS

Volterra integral equations have an integral with a variable limit. The Volterra equation of the second kind is

$$u(x) = f(x) + \lambda \int_a^x K(x, t)u(t)\, dt$$

whereas a Volterra equation of the first kind is

$$u(x) = \lambda \int_a^x K(x, t)u(t)\,dt$$

Equations of the first kind are very sensitive to solution errors so that they present severe numerical problems. Volterra equations are similar to initial value problems.

A Fredholm equation of the second kind is

$$u(x) = f(x) + \lambda \int_a^b K(x, t)u(t)\,dt$$

whereas a Fredholm equation of the first kind is

$$u(x) = \int_a^b K(x, t)u(t)\,dt$$

The limits of integration are fixed, and these problems are analogous to boundary value problems.

An eigenvalue problem is a homogeneous equation of the second kind, and solutions exist only for certain λ.

$$u(x) = \lambda \int_a^b K(x, t)u(t)\,dt$$

See Refs. 105 and 195 for further information and existence proofs.

If the unknown function u appears in the equation in any way except to the first power, the integral equation is said to be nonlinear. The equation $u(x) = f(x) + \int_a^b K(x, t)[u(t)]^{3/2}\,dt$ is nonlinear. The differential equation $du/dx = g(x, u)$ is equivalent to the nonlinear integral equation $u(x) = c + \int_a^x g[t, u(t)]\,dt$.

An integral equation is said to be singular when either one or both of the limits of integration become infinite or if $K(x, t)$ becomes infinite for one or more points of the interval under discussion.

Example $u(x) = x + \int_0^\infty \cos(xt)u(t)\,dt$ and $f(x) = \int_0^x \dfrac{u(t)}{x-t}\,dt$ are both singular. The kernel of the first equation is $\cos(xt)$, and that of the second is $(x-t)^{-1}$.

RELATION TO DIFFERENTIAL EQUATIONS

The Leibniz rule (see "Integral Calculus") can be used to show the equivalence of the initial-value problem consisting of the second-order differential equation $d^2y/dx^2 + A(x)(dy/dx) + B(x)y = f(x)$ together with the prescribed initial conditions $y(a) = y_0$, $y'(a) = y'_0$ to the integral equation.

$$y(x) = \int_a^x K(x, t)y(t)\,dt + F(x)$$

where $K(x, t) = (t - x)[B(t) - A'(t)] - A(t)$

and $F(x) = \int_a^x (x - t)f(t)\,dt + [A(a)y_0 + y'_0](x - a) + y_0$

This integral equation is a **Volterra equation** of the second kind. Thus the initial-value problem is equivalent to a Volterra integral equation of the second kind.

Example $d^2y/dx^2 + x^2(dy/dx) + xy = x$, $y(0) = 1$, $y'(0) = 0$. Here $A(x) = x^2$, $B(x) = x$, $f(x) = x$. The equivalent integral equation is $y(x) = \int_0^x K(x, t)y(t)\,dt + F(x)$ where $K(x, t) = t(x - t) - t^2$ and $F(x) = \int_0^x (x - t)t\,dt + 1 = x^3/6 + 1$. Combining these $y(x) = \int_0^x t[x - 2t]y(t)\,dt + x^3/6 + 1$.

Eigenvalue problems can also be related. For example, the problem $(d^2y/dx^2) + \lambda y = 0$ with $y(0) = 0$, $y(a) = 0$ is equivalent to the integral equation $y(x) = \lambda \int_0^a K(x, t)y(t)\,dt$, where $K(x, t) = (t/a)(a - x)$ when $t < x$ and $K(x, t) = (x/a)(a - t)$ when $t > x$. The differential equation may be recovered from the integral equation by differentiating the integral equation by using the Leibniz rule.

METHODS OF SOLUTION

In general, the solution of integral equations is not easy, and a few exact and approximate methods are given here. Often numerical methods must be employed, as discussed in "Numerical Solution of Integral Equations."

Equations of Convolution Type The equation $u(x) = f(x) + \lambda \int_0^x K(x - t)u(t)\,dt$ is a special case of the linear integral equation of the second kind of Volterra type. The integral part is the convolution integral discussed under "Integral Transforms (Operational Methods)"; so the solution can be accomplished by Laplace transforms; $L[u(x)] = L[f(x)] + \lambda L[u(x)]L[K(x)]$ or

$$L[u(x)] = \frac{L[f(x)]}{1 - \lambda L[K(x)]}, \quad u(x) = L^{-1}\left[\frac{L[f(x)]}{1 - \lambda L[K(x)]}\right]$$

Equations of the type considered here occur quite frequently in practice in what can be called "cause-and-effect" systems.

Example In a certain linear system, the effect $E(t)$ due to a cause $C = \lambda E$ at time τ is a function only of the elapsed time $t - \tau$. If the system has the activity level 1 at time $t < 0$, the cause λE and effect (E) relation is given by the integral equation $E(t) = 1 + \lambda \int_0^t K(t - \tau)E(\tau)\,d\tau$. Let $K(t - \tau) = t - \tau$. Then $E(t) = 1 + \lambda \int_0^t (t - \tau)E(\tau)\,d\tau$. By using the transform method

$$E(t) = L^{-1}\left[\frac{L[1]}{1 - \lambda L[K(t)]}\right] = L^{-1}\left[\frac{1/p}{1 - \lambda/p^2}\right] = L^{-1}\left[\frac{p}{p^2 - \lambda}\right] = \cosh\sqrt{\lambda}\,t$$

Method of Successive Approximations Consider the equation $y(x) = f(x) + \lambda \int_a^b K(x, t)y(t)\,dt$. In this method a unique solution is obtained in sequence form as follows: Substitute in the right-hand member of the equation $y_0(t)$ for $y(t)$. Upon integration there results $y_1(t) = f(x) + \lambda \int_a^b K(x, t)y_0(t)\,dt$. Continue in like manner by replacing y_0 by y_1, y_1 by y_2, etc. A series of functions $y_0(x)$, $y_1(x)$, $y_2(x)$, . . . are obtained which satisfy the equations

$$y_n(x) = f(x) + \lambda \int_a^b K(x, t)y_{n-1}(t)\,dt$$

Then $y_n(x) = f(x) + \lambda \int_a^b K(x, t)f(t)\,dt + \lambda^2 \int_a^b K(x, t) \int_a^b K(t, t_1)f(t_1)\,dt_1\,dt + \lambda^3 \int_a^b K(x, t) \int_a^b K(t, t_1) \int_a^b K(t_1, t_2)f(t_2)\,dt_2\,dt_1\,dt + \cdots + R_n$, where R_n is the remainder, and

$$|R_n| \le |\lambda^n| \begin{pmatrix} \text{max. } y_0 \\ a \le x \le b \end{pmatrix} M^n(b - a)^n$$

where M = maximum value of $|K|$ in the rectangle $a \le t \le b$, $a \le x \le b$. If $|\lambda|M(b - a) < 1$, $\lim_{n\to\infty} R_n = 0$. Then $y_n(x) \to y(x)$, which is the unique solution.

Example Consider the equation $y(x) = 1 + \lambda \int_0^1 (1 - 3xt)y(t)\,dt$.

$$y(x) = 1 + \lambda \int_0^1 (1 - 3xt)\,dt + \lambda^2 \int_0^1 (1 - 3xt) \int_0^1 (1 - 3tt_1)\,dt_1\,dt + \cdots$$

$$= 1 + \lambda\left(1 - \frac{3}{2}x\right) + \lambda^2\frac{1}{4} + \frac{1}{4}\lambda^3\left(1 - \frac{3}{2}x\right) + \frac{\lambda^4}{16} + \frac{1}{16}\lambda^5\left(1 - \frac{3}{2}x\right) + \cdots$$

$$= \left(1 + \frac{\lambda^2}{4} + \frac{\lambda^4}{16} + \cdots\right)\left(1 + \lambda\left(1 - \frac{3}{2}x\right)\right)$$

$$= \frac{1 + \lambda(1 - \frac{3}{2}x)}{1 - \frac{1}{4}\lambda^2}, \quad |\lambda| < 2$$

Example $dy/dx = x^2 + y$. $x_0 = 0$, $y_0 = 1$. This problem is equivalent to the integral equation $y = 1 + \int_0^x (x^2 + y)\,dx$. Let the initial approximation for y be 1. Then

$$y^{(1)} = 1 + \int_0^x (x^2 + 1)\,dx = 1 + x + \frac{x^3}{3}$$

$$y^{(2)} = 1 + \int_0^x [x^2 + y^{(1)}]\,dx = 1 + \int_0^x \left[x^2 + 1 + x + \frac{x^3}{3}\right]dx$$

$$= 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{12}, \text{ etc.}$$

INTEGRAL TRANSFORMS (OPERATIONAL METHODS)

REFERENCES: 63, 64, 71, 72, 97, 137, 217.

The term "operational method" implies a procedure of solving differential and difference equations by which the boundary or initial conditions are automatically satisfied in the course of the solution. The technique offers a very powerful tool in the applications of mathematics, but it is limited to linear problems.

Most integral transforms are special cases of the equation $g(s) = \int_a^b f(t)K(s, t)\,dt$ in which $g(s)$ is said to be the transform of $f(t)$ and $K(s, t)$ is called the kernel of the transform. A tabulation of the more important kernels and the interval (a, b) of applicability follows. The first three transforms are considered here.

Name of transform	(a, b)	$K(s, t)$
Laplace	$(0, \infty)$	e^{-st}
Fourier	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi}} e^{-ist}$
Fourier cosine	$(0, \infty)$	$\sqrt{\frac{2}{\pi}} \cos st$
Fourier sine	$(0, \infty)$	$\sqrt{\frac{2}{\pi}} \sin st$
Mellin	$(0, \infty)$	t^{s-1}
Hankel	$(0, \infty)$	$tJ_\nu(st)$, $\nu \geq -\frac{1}{2}$

LAPLACE TRANSFORM

The Laplace transform of a function $f(t)$ is defined by $F(s) = L\{f(t)\} = \int_0^\infty e^{-st}f(t)\,dt$, where s is a complex variable. Note that the transform is an improper integral and therefore may not exist for all continuous functions and all values of s. We restrict consideration to those values of s and those functions f for which this improper integral converges.

The function $L[f(t)] = g(s)$ is called the direct transform, and $L^{-1}[g(s)] = f(t)$ is called the inverse transform. Both the direct and the inverse transforms are tabulated for many often-occurring functions. In general,

$$L^{-1}[g(s)] = \frac{1}{2\pi i} \int_{\alpha - i\infty}^{\alpha + i\infty} e^{st} g(s)\,ds$$

and to evaluate this integral requires a knowledge of complex variables, the theory of residues, and contour integration.

A function is said to be piecewise continuous on an interval if it has only a finite number of finite (or jump) discontinuities. A function f on $0 < t < \infty$ is said to be of exponential growth at infinity if there exist constants M and α such that $|f(t)| \leq Me^{\alpha t}$ for sufficiently large t.

Sufficient Conditions for the Existence of Laplace Transform Suppose f is a function which is (1) piecewise continuous on every finite interval $0 < t < T$, (2) of exponential growth at infinity, and (3) $\int_0^\delta |f(t)|\,dt$ exist (finite) for every finite $\delta > 0$. Then the Laplace transform of f exists for all complex numbers s with sufficiently large real part.

Note that condition 3 is automatically satisfied if f is assumed to be piecewise continuous on every finite interval $0 \leq t < T$. The function $f(t) = t^{-1/2}$ is not piecewise continuous on $0 \leq t \leq T$ but satisfies conditions 1 to 3.

Let Λ denote the class of all functions on $0 < t < \infty$ which satisfy conditions 1 to 3.

Example Let $f(t)$ be the Heaviside step function at $t = t_0$; i.e., $f(t) = 0$ for $t \leq t_0$, and $f(t) = 1$ for $t > t_0$. Then

$$L\{f(t)\} = \int_{t_0}^\infty e^{-st}\,dt = \lim_{T\to\infty} \int_{t_0}^T e^{-st}\,dt = \lim_{T\to\infty} \frac{1}{s}\,(e^{-st_0} - e^{-sT}) = \frac{e^{-st_0}}{s}$$

provided $s > 0$.

Example Let $f(t) = e^{at}$, $t \geq 0$, where a is a real number. Then $L\{e^{at}\} = \int_0^\infty e^{-(s-a)}\,dt = 1/(s - a)$, provided Re $s > a$.

Properties of the Laplace Transform

1. The Laplace transform is a linear operator: $L\{af(t) + bg(t)\} = aL\{f(t)\} + bL\{g(t)\}$ for any constants a, b and any two functions f and g whose Laplace transforms exist.
2. The Laplace transform of a real-valued function is real for real s. If $f(t)$ is a complex-valued function, $f(t) = u(t) + iv(t)$, where u and v are real, then $L\{f(t)\} = L\{u(t)\} + iL\{v(t)\}$. Thus $L\{u(t)\}$ is the real part of $L\{f(t)\}$, and $L\{v(t)\}$ is the imaginary part of $L\{f(t)\}$.
3. The Laplace transform of a function in the class Λ has derivatives of all orders, and $L\{t^k f(t)\} = (-1)^k d^k F(s)/ds^k$, $k = 1, 2, 3, \ldots$.

Example $\int_0^\infty e^{-st} \sin at\,dt = \frac{a}{s^2 + a^2}$, $s > 0$. By property 3, $\frac{2as}{(s^2 + a^2)^2} = \int_0^\infty e^{-st}\, t \sin at\,dt = L\{t \sin at\}$.

Example By applying property 3 with $f(t) = 1$ and using the preceding results, we obtain

$$L\{t^k\} = (-1)^k \frac{d^k}{ds^k}\left(\frac{1}{s}\right) = \frac{k!}{s^{k+1}}$$

provided Re $s > 0$; $k = 1, 2, \ldots$. Similarly, we obtain

$$L\{t^k e^{at}\} = (-1)^k \frac{d^k}{ds^k}\left(\frac{1}{s - a}\right) = \frac{k!}{(s - a)^{k+1}}$$

4. Frequency-shift property (or, equivalently, the transform of an exponentially modulated function). If $F(s)$ is the Laplace transform of a function $f(t)$ in the class Λ, then for any constant a, $L\{e^{at}f(t)\} = F(s - a)$.

Example $L\{te^{-at}\} = \frac{1}{(s + a)^2}$, $s > 0$.

5. Time-shift property. Let $u(t - a)$ be the unit step function at $t = a$. Then $L\{f(t - a)u(t - a)\} = e^{-as}F(s)$.
6. Transform of a derivative. Let f be a differentiable function such that both f and f' belong to the class Λ. Then $L\{f'(t)\} = sF(s) - f(0)$.
7. Transform of a higher-order derivative. Let f be a function which has continuous derivatives up to order n on $(0, \infty)$, and suppose that f and its derivatives up to order n belong to the class Λ. Then $L\{f^{(j)}(t)\} = s^jF(s) - s^{j-1}f(0) - s^{j-2}f'(0) - \cdots - sf^{(j-2)}(0) - f^{(j-1)}(0)$ for $j = 1, 2, \ldots, k$.

Example $L\{f''(t)\} = s^2L\{f(t)\} - sf(0) - f'(0)$

$L\{f'''(t)\} = s^3L\{f(t)\} - s^2f(0) - sf'(0) - f''(0)$

Example Solve $y'' + y = 2e^t$, $y(0) = y'(0) = 2$. $L[y''] = -y'(0) - sy(0) + s^2L[y] = -2 - 2s + s^2L[y]$. Thus

$$-2 - 2s + s^2L[y] + L[y] = 2L[e^t] = \frac{2}{s - 1}$$

$$L[y] = \frac{2s^2}{(s - 1)(s^2 + 1)} = \frac{1}{s - 1} + \frac{s}{s^2 + 1} + \frac{1}{s^2 + 1}$$

Hence $y = e^t + \cos t + \sin t$.

A short table (Table 3-1) of very common Laplace transforms and inverse transforms follows. The references include more detailed tables. NOTE: $\Gamma(n + 1) = \int_0^\infty x^n e^{-x}\,dx$ (gamma function); $J_n(t)$ = Bessel function of the first kind of order n.

8. $L\left[\int_a^t f(t)\,dt\right] = \frac{1}{s}L[f(t)] + \frac{1}{s}\int_a^0 f(t)\,dt$

Example Find $f(t)$ if $L[f(t)] = \frac{1}{s^2}\left[\frac{1}{s^2 - a^2}\right]$ $\quad L\left[\frac{1}{a}\sinh at\right] = \frac{1}{s^2 - a^2}$

Therefore $f(t) = \int_0^t \left[\int_0^t \frac{1}{a}\sinh at\,dt\right] dt = \frac{1}{a^2}\left[\frac{\sinh at}{a} - t\right]$

TABLE 3-1 Laplace Transforms

$f(t)$	$g(s)$	$f(t)$	$g(s)$
1	$1/s$	$e^{-at}(1-at)$	$\dfrac{s}{(s+a)^2}$
t^n, $(n\ a + \text{integer})$	$\dfrac{n!}{s^{n+1}}$	$\dfrac{t \sin at}{2a}$	$\dfrac{s}{(s^2+a^2)^2}$
t^n, $n \neq +\text{integer}$	$\dfrac{\Gamma(n+1)}{s^{n+1}}$	$\dfrac{1}{2a^2} \sin at \sinh at$	$\dfrac{s}{s^4+4a^4}$
$\cos at$	$\dfrac{s}{s^2+a^2}$	$\cos at \cosh at$	$\dfrac{s^2}{s^4+4a^4}$
$\sin at$	$\dfrac{a}{s^2+a^2}$	$\dfrac{1}{2a}(\sinh at + \sin at)$	$\dfrac{s^2}{s^4-a^4}$
$\cosh at$	$\dfrac{s}{s^2-a^2}$	$\frac{1}{2}(\cosh at + \cos at)$	$\dfrac{s^2}{s^4-a^4}$
$\sinh at$	$\dfrac{a}{s^2-a^2}$	$\dfrac{\sin at}{t}$	$\tan^{-1}\dfrac{a}{s}$
e^{-at}	$\dfrac{1}{s+a}$	$J_0(at)$	$\dfrac{1}{\sqrt{s^2+a^2}}$
$e^{-bt}\cos at$	$\dfrac{s+b}{(s+b)^2+a^2}$	$na^n \dfrac{J_n(at)}{t}$	$\dfrac{1}{\sqrt{s^2+a^2}-s^n}$
$e^{-bt}\sin a$	$\dfrac{a}{(s+b)^2+a^2}$	$J_0(2\sqrt{at})$	$\dfrac{1}{s}e^{-a/s}$

9. $L\left[\dfrac{f(t)}{t}\right] = \int_s^\infty g(s)\,ds \qquad L\left[\dfrac{f(t)}{t^k}\right] = \underbrace{\int_s^\infty \cdots \int_s^\infty}_{k \text{ integrals}} g(s)(ds)^k$

Example $L\left[\dfrac{\sin at}{t}\right] = \int_s^\infty L[\sin at]\,ds = \int_s^\infty \dfrac{a\,ds}{s^2+a^2} = \cot^{-1}\dfrac{s}{a}$

10. The unit step function

$$u(t-a) = \begin{cases} 0 & t < a \\ 1 & t > a \end{cases} \qquad L[u(t-a)] = e^{-as}/s$$

11. The unit impulse function is

$$\delta(a) = u'(t-a) = \begin{cases} \infty & \text{at } t = a \\ 0 & \text{elsewhere} \end{cases} \qquad L[u'(t-a)] = e^{-as}$$

12. $L^{-1}[e^{-as}g(s)] = f(t-a)u(t-a)$ (second shift theorem).
13. If $f(t)$ is periodic of period b, i.e., $f(t+b) = f(t)$, then

$$L[f(t)] = \left[\frac{1}{1-e^{-bs}}\right]\int_0^b e^{-st} f(t)\,dt$$

Example The partial differential equations relating gas composition to position and time in a gas chromatograph are $\partial y/\partial n + \partial x/\partial\theta = 0$, $\partial y/\partial n = x - y$, where $x = mx'$, $n = (k_G aP/G_m)h$, $\theta = (mk_G aP/\rho_B)t$ and G_M = molar velocity, y = mole fraction of the component in the gas phase, ρ_B = bulk density, h = distance from the entrance, P = pressure, k_G = mass-transfer coefficient, and m = slope of the equilibrium line. These equations are equivalent to $\partial^2 y/\partial n\,\partial\theta + \partial y/\partial n + \partial y/\partial\theta = 0$, where the boundary conditions considered here are $y(0, \theta) = 0$ and $x(n, 0) = y(n, 0) + (\partial y/\partial n)\ (n, 0) = \delta(0)$ (see property 11). The problem is conveniently solved by using the Laplace transform of y with respect to n; write $g(s, \theta) = \int_0^\infty e^{-ns} y(n, \theta)\,dn$. Operating on the partial differential equation gives $s(dg/d\theta) - (\partial y/\partial\theta)\ (0, \theta) + sg - y(0, \theta) + dg/d\theta = 0$ or $(s+1)\ (dg/d\theta) + sg = (\partial y/\partial\theta)\ (0, \theta) + y(0, \theta) = 0$. The second boundary condition gives $g(s, 0) + sg(s, 0) - y(0, 0) = 1$ or $g(s, 0) + sg(s, 0) = 1$ ($L[\delta(0)] = 1$). A solution of the ordinary differential equation for g consistent with this second condition is

$$g(s, \theta) = \frac{1}{s+1} e^{-s\theta/(s+1)}$$

Inversion of this transform gives the solution $y(n, \theta) = e^{-(n+\theta)}\ I_0(2\sqrt{n\theta})$ where I_0 = zero-order Bessel function of an imaginary argument. For large u, $I_n(u) \sim e^u/\sqrt{2\pi u}$. Hence for large n,

$$y(n, \theta) \sim \frac{\exp\left[-(\sqrt{\theta} - \sqrt{n})^2\right]}{2\pi^{1/2}(n\theta)^{1/4}}$$

or for sufficiently large n, the peak concentration occurs near $\theta = n$.

Other applications of Laplace transforms are given under "Differential Equations."

CONVOLUTION INTEGRAL

The convolution integral (faltung) of two functions $f(t)$, $r(t)$ is $x(t) = f(t)\circ r(t) = \int_0^t f(\tau) r(t-\tau)\,d\tau$.

Example $t \circ \sin t = \int_0^t \tau \sin(t-\tau)\,d\tau = t - \sin t.$

$$L[f(t)]L[h(t)] = L[f(t)\circ h(t)]$$

z-TRANSFORM

See Refs. 198, 218, and 256. The z-transform is useful when data is available at only discrete points. Let

$$f^*(t) = f(t_k)$$

be the value of f at the sample points

$$t_k = k\,\Delta t, \quad k = 0, 1, 2, \ldots$$

Then the function $f^*(t)$ is

$$f^*(t) = \sum_{k=0}^{\infty} f(t_k)\ \delta(t - t_k)$$

Take the Laplace transform of this.

$$g^*(s) = L[f^*(t)] = \sum_{k=0}^{\infty} f(t_k)\ e^{-stk} = \sum_{k=0}^{\infty} f(t_k)\ e^{-s\Delta tk}$$

For convenience, replace $e^{-s\Delta t}$ by z and call $g^*(z)$ the z-transform of $f^*(t)$.

$$g^*(z) = \sum_{k=0}^{\infty} f(t_k)\ z^{-k}$$

The z-transform is used in process control when the signals are at intervals of Δt. A brief table (Table 3-2) is provided here.

The z-transform can also be used to solve difference equations, just like the Laplace transform can be used to solve differential equations.

TABLE 3-2 z-Transforms

$f(k)$	$g^*(z)$
$1(k)$	$\dfrac{1}{1-z^{-1}}$
$k\,\Delta t$	$\dfrac{\Delta t\, z^{-1}}{(1-z^{-1})^2}$
$(k\,\Delta t)^{n-1}$	$\lim_{a\to 0}(-1)^{n-1}\dfrac{\partial^{n-1}}{\partial a^{n-1}}\left(\dfrac{1}{1-e^{-a\Delta t}z^{-1}}\right)$
$\sin a\,k\,\Delta t$	$\dfrac{z^{-1}\sin a\,\Delta t}{(1-2z^{-1}\cos a\,\Delta t + z^{-2})}$
$\cos a\,k\,\Delta t$	$\dfrac{1-z^{-1}\cos a\,\Delta t}{(1-2z^{-1}\cos a\,\Delta t + z^{-2})}$
$e^{-ak\Delta t}$	$\dfrac{1}{1-e^{-a\Delta t}z^{-1}}$
$e^{-bk\Delta t}\cos a\,k\,\Delta t$	$\dfrac{1-z^{-1}e^{-b\Delta t}\cos a\,\Delta t}{1-2z^{-1}e^{-b\Delta t}\cos a\,\Delta t + z^{-2}e^{-2b\Delta t}}$
$\dfrac{1}{b}e^{-bk\Delta t}\sin a\,k\,\Delta t$	$\dfrac{1}{b}\dfrac{z^{-1}e^{-b\Delta t}\sin a\,\Delta t}{1-2z^{-1}e^{-b\Delta t}\cos a\,\Delta t + z^{-2}e^{-2b\Delta t}}$

Example The difference equation for $y(k)$ is

$$y(k) + a_1 y(k-1) + a_2 y(k-2) = b_1 u(k)$$

Take the z-transform

$$(1 + a_1 z^{-1} + a_2 z^{-2})\, y^*(z) = u^*(z)$$

Then

$$y^*(z) = \frac{u^*(z)}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

The inverse transform must be found, usually from a table of inverse transforms.

FOURIER TRANSFORM

The Fourier transform is given by

$$F[f(t)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-ist}\, dt = g(s)$$

and its inverse by

$$F^{-1}[g(s)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(s) e^{ist}\, dt = f(t)$$

In brief, the condition for the Fourier transform to exist is that $\int_{-\infty}^{\infty} |f(t)|\, dt < \infty$, although certain functions may have a Fourier transform even if this is violated.

Example The function $f(t) = \begin{cases} 1 & -a \le t \le a \\ 0 & \text{elsewhere} \end{cases}$ has $F[f(t)] = \int_{-a}^{a} e^{-ist}\, dt = \int_0^a e^{ist}\, dt + \int_0^a e^{-ist}\, dt = 2\int_0^a \cos st\, dt = \dfrac{2 \sin sa}{s}$

Properties of the Fourier Transform Let $F[f(t)] = g(s)$; $F^{-1}[g(s)] = f(t)$.

1. $F[f^{(n)}(t)] = (is)^n F[f(t)]$.
2. $F[af(t) + bh(t)] = aF[f(t)] + bF[h(t)]$.
3. $F[f(-t)] = g(-s)$.
4. $F[f(at)] = \dfrac{1}{a} g\left(\dfrac{s}{a}\right), a > 0.$
5. $F[e^{-iwt} f(t)] = g(s + w)$.
6. $F[f(t + t_1)] = e^{ist_1} g(s)$.
7. $F[f(t)] = G(is) + G(-is)$ if $f(t) = f(-t)$ (f even)
   $F[f(t)] = G(is) - G(-is)$ if $f(t) = -f(-t)$ (f odd)

where $G(s) = L[f(t)]$. This result allows the use of the Laplace-transform tables to obtain the Fourier transforms.

Example Find $F[e^{-a|t|}]$ by property 7. $e^{-a|t|}$ is even. So $L[e^{-at}] = 1/(s + a)$. Therefore, $F[e^{-a|t|}] = 1/(is + a) + 1/(-is + a) = 2a/(s^2 + a^2)$.

Tables of this transform may be found in *Higher Transcendental Functions,* vols. I, II, and III, A. Erdelyi, et al., McGraw-Hill, New York, 1953–1955.

FOURIER COSINE TRANSFORM

The Fourier cosine transform is given by

$$F_c[f(t)] = g(s) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(t) \cos st\, dt$$

and its inverse by

$$F_c^{-1}[g(s)] = f(t) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} g(s) \cos st\, ds$$

The Fourier sine transform F_s is obtainable by replacing the cosine by the sine in these integrals.

Example $F_c[f(t)]$, $f(t) = \begin{cases} 1 & 0 < t < a \\ 0 & a < t < \infty \end{cases}$ $F_c[f(t)] = \sqrt{\dfrac{2}{\pi}} \int_0^a \cos st\, dt = \sqrt{\dfrac{2}{\pi}} \dfrac{\sin as}{s}$

Properties of the Fourier Cosine Transform $F_c[f(t)] = g(s)$.

1. $F_c[af(t) + bh(t)] = aF_c[f(t)] + bF_c[h(t)]$.
2. $F_c[f(at)] = (1/a) g(s/a)$.
3. $F_c[f(at) \cos bt] = \dfrac{1}{2a}\left[g\left(\dfrac{s+b}{a}\right) + g\left(\dfrac{s-b}{a}\right)\right], a, b > 0$
4. $F_c[t^{-2n} f(t)] = (-1)^n (d^{2n} g/ds^{2n})$.
5. $F_c[t^{2n+1} f(t)] = (-1)^n (d^{2n+1}/ds^{2n+1})\, F_s[f(t)]$.

A short table (Table 3-3) of Fourier cosine transforms follows.

TABLE 3-3 Fourier Transforms

$f(t)$		$\dfrac{g(s)}{\sqrt{2/\pi}}$
t $2 - t$ 0	$0 < t < 1$ $1 < t < 2$ $2 < t < \infty$	$\dfrac{1}{s^2}[2 \cos s - 1 - \cos 2s]$
$t^{-1/2}$		$\pi^{1/2}(2s)^{-1/2}$
0 $(t-a)^{-1/2}$	$0 < t < a$ $a < t < \infty$	$\pi^{1/2}(2s)^{-1/2}[\cos as - \sin as]$
$(t^2 + a^2)^{-1}$		$\frac{1}{2}\pi a^{-1} e^{-as}$
e^{-at}	$a > 0$	$\dfrac{a}{s^2 + a^2}$
e^{-at^2}	$a > 0$	$\frac{1}{2}\pi^{1/2} a^{-1/2} e^{-s^2/4a}$
$\dfrac{\sin at}{t}$	$a > 0$	$\begin{cases} \pi/2 & s < a \\ \pi/4 & s = a \\ 0 & s > a \end{cases}$

Example The temperature θ in the semi-infinite rod $0 \le x < \infty$ is determined by the differential equation $\partial\theta/\partial t = k(\partial^2\theta/\partial x^2)$ and the condition $\theta = 0$ when $t = 0$, $x \ge 0$; $\partial\theta/\partial x = -\mu$ = constant when $x = 0$, $t > 0$. By using the Fourier cosine transform a solution may be found as

$$\theta(x, t) = \frac{2\mu}{\pi} \int_0^{\infty} \frac{\cos px}{p} (1 - e^{-kp^2 t})\, dp.$$

MATRIX ALGEBRA AND MATRIX COMPUTATIONS

REFERENCES: General (textbooks that cover at an introductory level a variety of topics that constitute a core of numerical methods for practicing engineers), 2, 3, 4, 22, 56, 59, 70, 77, 133, 135, 143, 150, 155, 219. Numerical solution of nonlinear equations, 153, 171, 237, 302. Numerical solution of ordinary differential equations, 76, 117, 127, 185, 257. Numerical solution of integral equations, 23, 26, 129, 162. Numerical solution of partial differential equations, 11, 76, 127, 133, 155, 210, 251, 286, 287, 213, 233, 253. Spline functions and applications, 38, 56, 70, 230. Finite elements and applications, 5, 29, 83, 130, 164, 210, 241, 281, 287, 303, 304. Fast Fourier transforms, 47, 56, 135, 238. Software, 187, 231.

MATRIX ALGEBRA

Matrices A rectangular array of mn quantities, arranged in m rows and n columns

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

is called a matrix. The elements a_{ij} may be real or complex. The notation a_{ij} means the element in the ith row and jth column, i is called the row index, j the column index. If $m = n$ the matrix is said to be square and of order n. A matrix, even if it is square, does not have a numerical value, as a determinant does. However, if the matrix A is square, a determinant can be formed which has the same elements as the matrix A. This is called the determinant of the matrix and is written det (A) or $|A|$. If A is square and det $(A) \neq 0$, A is said to be nonsingular; if det $(A) = 0$, A is said to be singular. A matrix A has rank r if and only if it has a nonvanishing determinant of order r and no nonvanishing determinant of order $> r$.

Equality of Matrices Let $A = (a_{ij})$, $B = (b_{ij})$. Two matrices A and B are *equal* $(=)$ if and only if they are identical; that is, they have the same number of rows and the same number of columns and equal corresponding elements $(a_{ij} = b_{ij}$ for all i and $j)$.

Addition and Subtraction The operations of addition $(+)$ and subtraction $(-)$ of two or more matrices are possible if and only if they have the same number of rows and columns. Thus $A \pm B = (a_{ij} \pm b_{ij})$; i.e., addition and subtraction are of corresponding elements.

Transposition The matrix obtained from A by interchanging the rows and columns of A is called the transpose of A, written A' or A^T.

Example $A = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 1 & 6 \end{bmatrix}$, $A^T = \begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 4 & 6 \end{bmatrix}$

Note that $(A^T)^T = A$.

Multiplication Let $A = (a_{ij})$, $i = 1, \ldots, m_1; j = 1, \ldots, m_2$. $B = (b_{ij})$, $i = 1, \ldots, n_1, j = 1, \ldots, n_2$. The product AB is defined if and only if the number of columns of A (m_2) equals the number of rows of $B(n_1)$, i.e., $n_1 = m_2$. For two such matrices the product $P = AB$ is defined by summing the element by element products of a row of A by a column of B.

This is the row by column rule. Thus

$$p_{ij} = \sum_{k=1}^{n_1} a_{ik} b_{kj}$$

The resulting matrix has m_1 rows and n_2 columns.

Example $\begin{bmatrix} 3 & 2 \\ 1 & 1 \\ 5 & 4 \end{bmatrix} \begin{bmatrix} 0 & 1 & 5 & 6 \\ -2 & 0 & 1 & 3 \end{bmatrix} = \begin{bmatrix} -4 & 3 & 17 & 24 \\ -2 & 1 & 6 & 9 \\ -8 & 5 & 29 & 42 \end{bmatrix}$

It is helpful to remember that the element p_{ij} is formed from the ith row of the first matrix and the jth column of the second matrix. The matrix product is not commutative. That is, $AB \neq BA$ in general.

Inverse of a Matrix A square matrix A is said to have an inverse if there exists a matrix B such that $AB = BA = I$, where I is the identity matrix of order n.

$$\begin{bmatrix} 1 & 0 \ldots . & 0 \\ 0 & 1 \ldots & \\ \vdots & & 1 \quad 0 \\ 0 & \ldots . 0 & 1 \end{bmatrix}$$

The inverse B is a square matrix of the order of A, designated by A^{-1}. Thus $AA^{-1} = A^{-1}A = I$. A square matrix A has an inverse if and only if A is nonsingular.

Certain relations are important:

(1) $(AB)^{-1} = B^{-1}A^{-1}$

(2) $(AB)^T = B^T A^T$

(3) $(A^{-1})^T = (A^T)^{-1}$

(4) $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$

Scalar Multiplication Let c be any real or complex number. Then $cA = (ca_{ij})$.

Adjugate Matrix of a Matrix Let A_{ij} denote the cofactor of the element a_{ij} in the determinant of the matrix A. The matrix B^T where $B = (A_{ij})$ is called the adjugate matrix of A written adj $A = B^T$. The elements b_{ij} are calculated by taking the matrix A, deleting the ith row and jth column, and calculating the determinant of the remaining matrix times $(-1)^{i+j}$. Then $A^{-1} = \text{adj } A/|A|$. This definition may be used to calculate A^{-1}. However, it is very laborious and the inversion is usually accomplished by numerical techniques shown under "Numerical Analysis and Approximate Methods."

Example Let $A = \begin{bmatrix} 3 & 0 & -1 \\ -1 & 2 & 1 \\ 3 & 6 & 3 \end{bmatrix}$ Form $B = (A_{ij})$, $B = \begin{bmatrix} 0 & 6 & -12 \\ -6 & 12 & -18 \\ 2 & -2 & 6 \end{bmatrix}$

$$\text{adj } A = B^T = \begin{bmatrix} 0 & -6 & 2 \\ 6 & 12 & -2 \\ -12 & -18 & 6 \end{bmatrix}; \quad |A| = 12$$

$$A^{-1} = \frac{\text{adj } A}{|A|} = \begin{bmatrix} 0 & -\frac{1}{2} & \frac{1}{6} \\ \frac{1}{2} & 1 & -\frac{1}{6} \\ -1 & -\frac{3}{2} & \frac{1}{2} \end{bmatrix}$$

Linear Equations in Matrix Form Every set of n nonhomogeneous linear equations in n unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad\qquad & \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

can be written in matrix form as $AX = B$, where $A = (a_{ij})$, $X^T = [x_1 \cdots x_n]$, and $B^T = [b_1 \cdots b_n]$. The solution for the unknowns is $X = A^{-1}B$.

Special Square Matrices

1. A triangular matrix is a matrix all of whose elements above or below the main diagonal (set of elements $a_{11}, \ldots, a_{nn}$) are zero.

If A is triangular, det $(A) = a_{11} \cdot a_{22} \ldots a_{nn}$.

2. A diagonal matrix is one such that all elements both above and below the main diagonal are zero (i.e., $a_{ij} = 0$ for all $i \neq j$). If all diagonal elements are equal, the matrix is called scalar. If A is diagonal, $A = (a_{ij})$, $A^{-1} = (1/a_{ij})$.
3. If $a_{ij} = a_{ji}$ for all i and j (i.e., $A = A^T$), the matrix is symmetric.
4. If $a_{ij} = -a_{ji}$ for $i \neq j$ but the a_{ij} are not all zero, the matrix is skew.
5. If $a_{ij} = -a_{ji}$ for all i and j (i.e., $a_{ii} = 0$), the matrix is skew symmetric.
6. If $A^T = A^{-1}$, the matrix A is orthogonal.
7. If the matrix $A^* = (\bar{a}_{ij})^T$, $\bar{a}_{ij}$ = complex conjugate of a_{ij}, A^* is the hermitian conjugate of A.
8. If $A = A^{-1}$, A is involutory.
9. If $A = A^*$, A is hermitian.
10. If $A = -A^*$, A is skew hermitian.
11. If $A^{-1} = A^*$, A is unitary.

If A is any matrix, then AA^T and A^TA are square symmetric matrices, usually of different order.

Example Let $A = \begin{bmatrix} 5 & 1 & 3 & 0 \\ 3 & 4 & 1 & 5 \\ 2 & -2 & 0 & 1 \end{bmatrix}$, $A^T = \begin{bmatrix} 5 & 3 & 2 \\ 1 & 4 & -2 \\ 3 & 1 & 0 \\ 0 & 5 & 1 \end{bmatrix}$

$$AA^T = \begin{bmatrix} 35 & 22 & 8 \\ 22 & 51 & 3 \\ 8 & 3 & 9 \end{bmatrix}, \quad A^TA = \begin{bmatrix} 38 & 13 & 18 & 17 \\ 13 & 21 & 7 & 18 \\ 18 & 7 & 10 & 5 \\ 17 & 18 & 5 & 26 \end{bmatrix}$$

Using a program such as MATLAB, these are easily calculated.

Matrix Calculus

Differentiation Let the elements of $A = [a_{ij}(t)]$ be differentiable functions of t. Then $\dfrac{dA}{dt} = \left[\dfrac{da_{ij}(t)}{dt}\right]$.

Example $A=\begin{bmatrix}\sin t & \cos t\\ -\cos t & \sin t\end{bmatrix}$, $\dfrac{dA}{dt}=\begin{bmatrix}\cos t & -\sin t\\ \sin t & \cos t\end{bmatrix}$.

Integration The integral $\int A\,dt=[\int a_{ij}(t)\,dt]$.

Example $A=\begin{bmatrix}t & 2\\ t^2 & e^t\end{bmatrix}$, $\int A\,dt=\begin{bmatrix}t^2/2 & 2t\\ t^3/3 & e^t\end{bmatrix}$.

The matrix $B=A-\lambda I$ is called the characteristic (eigen) matrix of A. Here A is square of order n, λ is a scalar parameter, and I is the $n\times n$ identity. $\det B=\det(A-\lambda I)=0$ is the characteristic (eigen) equation for A. The characteristic equation is always of the same degree as the order of A. The roots of the characteristic equation are called the eigenvalues of A.

Example $A=\begin{bmatrix}1 & 2\\ 3 & 8\end{bmatrix}$, $B=\begin{bmatrix}1 & 2\\ 3 & 8\end{bmatrix}-\begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}=\begin{bmatrix}1-\lambda & 2\\ 3 & 8-\lambda\end{bmatrix}$

is the characteristic matrix and $f(\lambda)=\det(B)=\det(A-\lambda I)=(1-\lambda)(8-\lambda)-6=2-9\lambda+\lambda^2=0$ is the characteristic equation. The eigenvalues of A are the roots of $\lambda^2-9\lambda+2=0$, which are $(9\pm\sqrt{73})/2$.

A nonzero matrix X_i, which has one column and n rows, called a column vector satisfying the equation

$$(A-\lambda I)X_i=0$$

and associated with the ith characteristic root λ_i is called an eigenvector.

Vector and Matrix Norms To carry out error analysis for approximate and iterative methods for the solutions of linear systems, one needs notions for vectors in R^n and for matrices that are analogous to the notion of length of a geometric vector. Let R^n denote the set of all vectors with n components, $x=(x_1,\ldots,x_n)$. In dealing with matrices it is convenient to treat vectors in R^n as columns, and so $x=(x_1,\ldots,x_n)^T$; however, we shall here write them simply as row vectors. A norm on R^n is a real-valued function f defined on R^n with the following properties:

1. $f(x)\ge 0$ for all $x\in R^n$.
2. $f(x)=0$ if and only if $x=(0,0,\ldots,0)$.
3. $f(ax)=|a|f(x)$ for all real numbers a and $x\in R^n$.
4. $f(x+y)\leqq f(x)+f(y)$ for all $x, y\in R^n$.

The usual notation for a norm is $f(x)=\|x\|$.

The norm of a matrix is

$$\kappa(A)\equiv\|A\|\,\|A^{-1}\|$$

where $$\|A\|=\sup_{x\neq 0}\frac{\|Ax\|}{\|x\|}=\max_k\sum_{j=1}^{n}|a_{jk}|$$

The norm is useful when doing numerical calculations. If the computer's floating-point precision is 10^{-6}, then $\kappa=10^6$ indicates an ill-conditioned matrix. If the floating-point precision is 10^{-12} (double precision), then a matrix with $\kappa=10^{12}$ may be ill-conditioned. Two other measures are useful and are more easily calculated:

$$\text{Ratio}=\frac{\max_k|a_{kk}^{(k)}|}{\min_k|a_{kk}^{(k)}|},\quad V=\frac{|\det A|}{\alpha_1\,\alpha_2\ldots\alpha_n},\quad \alpha_i=(a_{i1}^2+a_{i2}^2+\cdots+a_{in}^2)^{1/2}$$

where $a_{kk}^{(k)}$ are the diagonal elements of the LU decomposition.

MATRIX COMPUTATIONS

The principal topics in linear algebra involve systems of linear equations, matrices, vector spaces, linear transformations, eigenvalues and eigenvectors, and least-squares problems. The calculations are routinely done on a computer.

***LU* Factorization of a Matrix** To every $m\times n$ matrix A there exists a permutation matrix P, a lower triangular matrix L with unit diagonal elements, and an $m\times n$ (upper triangular) echelon matrix U such that $PA=LU$. The Gauss elimination is in essence an algorithm to determine U, P, and L. The permutation matrix P may be needed since it may be necessary in carrying out the Gauss elimination to interchange two rows of A to produce a (nonzero) pivot, such as if we start with

$$A=\begin{bmatrix}0 & 2\\ 1 & 6\end{bmatrix}$$

If A is a square matrix and if principal submatrices of A are all nonsingular, then we may choose P as the identity in the preceding factorization and obtain $A=LU$. This factorization is unique if L is normalized (as assumed previously), so that it has unit elements on the main diagonal.

Solution of *Ax* = *b* by Using *LU* Factorization Suppose that the indicated system is compatible and that $A=LU$ (the case $PA=LU$ is similarly handled and amounts to rearranging the equations). Let $z=Ux$. Then $Ax=LUx=b$ implies that $Lz=b$. Thus to solve $Ax=b$ we first solve $Lz=b$ for z and then solve $Ux=z$ for x. This procedure does not require that A be invertible and can be used to determine all solutions of a compatible system $Ax=b$. Note that the systems $Lz=b$ and $Ux=z$ are both in triangular forms and thus can be easily solved.

The LU decomposition is essentially a Gaussian elimination, arranged for maximum efficiency (Ref. 112). The chief reason for doing an LU decomposition is that it takes fewer multiplications than would be needed to find an inverse. Also, once the LU decomposition has been found, it is possible to solve for multiple right-hand sides with little increase in work. The multiplication count for an $n\times n$ matrix and m right-hand sides is

$$\text{operation count}=\frac{1}{3}n^3-\frac{1}{3}n+mn^2$$

If an inverse is desired, it can be calculated by solving for the LU decomposition and then solving n problems with right-hand sides consisting of all zeroes except one entry. Thus $4n^2/3-n/3$ multiplications are required for the inverse. The determinant is given by

$$\text{Det }\mathbf{A}=\prod_{i=1}^{n}a_{ii}^{(i)}$$

where $a_{ii}^{(i)}$ are the diagonal elements obtained in the LU decomposition.

A tridiagonal matrix is one in which the only nonzero entries lie on the main diagonal and the diagonal just above and just below the main diagonal. The set of equations can be written as

$$a_ix_{i-1}+b_ix_i+c_ix_{i+1}=d_i$$

The LU decomposition is

$b_1=b_1$
for k=2,n do
$\quad a_k'=\dfrac{a_k}{b_{k-1}'},\quad b_k'=b_k-\dfrac{a_k}{b_{k-1}'}c_{k-1}$
enddo
$d_1'=d_1$
for k=2,n do
$\quad d_k'=d_k-a_k'\,d_{k-1}'$
enddo
$x_n=d_n'/b_n'$
for k=n−1,1 do
$\quad x_k=\dfrac{d_k'-c_k\,x_{k+1}}{b_k'}$
enddo

The operation count for an $n\times n$ matrix with m right-hand sides is

$$2(n-1)+m(3n-2)$$

If $$|b_i|>|a_i|+|c_i|$$

no pivoting is necessary, and this is true for many boundary-value problems and partial-differential equations.

Sparse matrices are ones in which the majority of the elements are

zero. If the structure of the matrix is exploited, the solution time on a computer is greatly reduced. See Refs. 27, 55, 95, 96, 101, and 246. The conjugate gradient method is one method for solving sparse matrix problems, since it only involves multiplication of a matrix times a vector. Thus the sparseness of the matrix is easy to exploit. The conjugate gradient method is an iterative method that converges for sure in n iterations where the matrix is an $n \times n$ matrix. See Refs. 142 and 206. The singular value decomposition is useful when the matrix is singular or nearly singular (see Ref. 231).

Matrix methods, in particular finding the rank of the matrix, can be used to find the number of independent reactions in a reaction set. If the stoichiometric numbers for the reactions and molecules are put in the form of a matrix, the rank of the matrix gives the number of independent reactions (see Ref. 13).

Pivoting in Gauss Elimination It might seem that the Gauss elimination completely disposes of the problem of finding solutions of linear systems, and theoretically it does. In practice, however, things are not so simple.

Example Assume three-decimal floating arithmetic (i.e., only the three most significant digits of any number are retained), and solve the following system by Gauss elimination:

$$0.000100x_1 + 1.00x_2 = 1.00$$
$$1.00x_1 + 1.00x_2 = 2.00$$

We obtain

$$0.100 \times 10^{-3}x_1 + 0.100 \times 10^{1}x_2 = 0.100 \times 10^{1}$$
$$-0.100 \times 10^{5}x_2 = -0.100 \times 10^{5}$$

so that $x_2 = 1.00$ and $x_1 = 0.00$.

We check our solution by computing the residual vector $\mathbf{r} = b - Ax$:

$$r_1 = 0.100 \times 10^{1} - 0.100 \times 10^{-3}x_1 - 0.100 \times 10^{1}x_2 = 0.00$$
$$r_2 = 0.200 \times 10^{1} - 0.100 \times 10^{1}x_1 - 0.100 \times 10^{1}x_2 = 0.100 \times 10^{1}$$

The fact that $r_2 = 1$ indicates that our "solution" is not very good. Indeed the exact solution of the system is $x_1 = 1.00010$ and $x_2 = 0.99990$, so the result computed by Gauss elimination is pretty bad.

Now reverse the order of the equations (that is, pivot) and solve

$$0.100 \times 10^{1}x_1 + 0.100 \times 10^{1}x_2 = 0.200 \times 10^{1}$$
$$0.100 \times 10^{1}x_2 = 0.100 \times 10^{1}$$

so that $x_2 = 1.00$ and $x_1 = 1.00$. In this case the residual vector is $r_1 = 0.00$ and $r_2 = 0.100 \times 10^{-3}$, a considerable improvement over the previous result. In fact, the solution is as good as one could hope for by using three-digit arithmetic.

The moral of the preceding example is that the order of equations can make a large difference in how good an answer is obtained. It should be clear that the poor results in the first case are caused by having the large multiplier $(0.100 \times 10^{1})/(0.100 \times 10^{-3})$, which resulted from dividing by a relatively small a_{11}. It is not enough just to avoid zero "pivots"; one must also avoid using pivots that are relatively small.

This magnification of errors can be reduced if we arrange that the pivot at any stage is larger in magnitude than any remaining element in the column. If this is done, the multipliers will then be less than or equal to 1 in magnitude. Gauss elimination modified in this manner is called pivotal condensation or partial pivoting. This is routinely done by computer programs.

NUMERICAL APPROXIMATIONS TO SOME EXPRESSIONS

APPROXIMATION IDENTITIES

For the following relationships the sign $\cong$ means approximately equal to, when X is small:

Approximation	Approximation
$\dfrac{1}{1 \pm X} \cong 1 \mp X$	$\sqrt{1 \pm X} \cong 1 \pm \dfrac{X}{2}$
$\dfrac{1+Y}{1 \mp X} \cong 1 + Y \pm X$	$(1 \pm X)^{-n} \cong 1 \mp nX$
$(1 \pm X)^{n} \cong 1 \pm nX$	$(1 \pm X)^{-1/2} \cong 1 \mp \dfrac{X}{2}$
$(a \pm X)^2 \cong a^2 \pm 2aX$	$e^{z} \cong 1 + X$
$\sin X \cong X (X \text{ rad})$	$\tan X \cong X$
$\sqrt{Y(Y+X)} \cong \dfrac{2Y+X}{2}$	$\sqrt{Y^2+X^2} \cong Y + \dfrac{X^2}{2Y}\left(\dfrac{X}{Y} \text{ small}\right)$

NUMERICAL ANALYSIS AND APPROXIMATE METHODS

REFERENCES: General (textbooks that cover at an introductory level a variety of topics that constitute a core of numerical methods for practicing engineers), 2, 3, 4, 22, 56, 59, 70, 77, 133, 135, 143, 150, 155, 219. Numerical solution of nonlinear equations, 153, 171, 237, 302. Numerical solution of ordinary differential equations, 76, 117, 127, 185, 257. Numerical solution of integral equations, 23, 26, 129, 162. Numerical solution of partial differential equations, 11, 76, 127, 133, 155, 210, 251, 286, 287, 213, 233, 253. Spline functions and applications, 38, 56, 70, 230. Finite elements and applications, 5, 29, 83, 130, 164, 210, 241, 281, 287, 303, 304. Fast Fourier transforms, 47, 56, 135, 238. Software, 187, 231.

INTRODUCTION

The goal of approximate and numerical methods is to provide convenient techniques for obtaining useful information from mathematical formulations of physical problems. Often this mathematical statement is not solvable by analytical means. Or perhaps analytic solutions are available but in a form that is inconvenient for direct interpretation numerically. In the first case it is necessary either to attempt to approximate the problem satisfactorily by one which will be amenable to analysis, to obtain an approximate solution to the original problem by numerical means, or to use the two techniques in combination.

Numerical techniques therefore do not yield exact results in the sense of the mathematician. Since most numerical calculations are inexact, the concept of error is an important feature. The error associated with an approximate value is defined as

$$\text{True value} = \text{approximate value} + \text{error}$$

The four sources of error are as follows:

1. *Gross errors.* These result from unpredictable human, mechanical, or electrical mistakes.
2. *Round-off errors.* These are the consequence of using a number specified by m correct digits to approximate a number which requires more than m digits for its exact specification. For example, approximate the irrational number $\sqrt{2}$ by 1.414. Such errors are often

present in experimental data, in which case they may be called inherent errors, due either to empiricism or to the fact that the computer dictates the number of digits. Such errors may be especially damaging in areas such as matrix inversion or the numerical solution of partial differential equations when the number of algebraic operations is extremely large.

3. *Truncation errors.* These errors arise from the substitution of a finite number of steps for an infinite sequence of steps which would yield the exact result. To illustrate this error consider the infinite series for $e^{-x} \cdot e^{-x} = 1 - x + x^2/2 - x^3/6 + E_T(x)$, where E_T is the truncation error, $E_T = (1/24)e^{-\varepsilon}x^4$, $0 < \varepsilon < x$. If x is positive, ε is also positive. Hence $e^{-\varepsilon} < 1$. The approximation $e^{-x} \approx 1 - x + x^2/2 - x^3/6$ is in error by a positive amount smaller than $(1/24)x^4$.

4. *Inherited errors.* These arise as a result of errors occurring in the previous steps of the computational algorithm.

The study of errors in a computation is related to the theory of probability. In what follows a relation for the error will be given in certain instances.

NUMERICAL SOLUTION OF LINEAR EQUATIONS

See the section entitled "Matrix Algebra and Matrix Computation."

NUMERICAL SOLUTION OF NONLINEAR EQUATIONS IN ONE VARIABLE

Special Methods for Polynomials Consider a polynomial equation of degree n:

$$P(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n = 0 \qquad (3\text{-}71)$$

with real coefficients. $P(x)$ has exactly n roots, which may be real or complex. If all the coefficients of $P(x)$ are integers, then any rational root, say, r/s (r, s integers, having no common divisors) of $P(x)$, must be such that r is an integral divisor of a_n and s is an integral divisor of a_0. Further, any polynomial with rational coefficients may be converted into one with integral coefficients by multiplying by the lowest common multiple of the denominators of the coefficients.

Example $3x^4 - \frac{5}{3}x^2 + \frac{1}{3}x - 2 = 0$. The lowest common multiple of the denominator is 15. Thus multiplying by 15 (which does not change the roots) gives $45x^4 - 25x^2 + 3x - 30 = 0$. The only possible rational roots r/s are such that r may have the values $\pm30, \pm15, \pm10, \pm6, \pm5, \pm3, \pm2, \pm1$. s may have the values $\pm45, \pm15, \pm9, \pm5, \pm3, \pm1$. The possible rational roots may then be formed from all possible quotients, having no common factor.

In addition to these results, one can obtain an upper and lower bound for the real roots by the following device: If $a_0 > 0$ in Eq. (3-71) and if in Eq. (3-71) the first negative coefficient is preceded by k coefficients which are positive or zero, and if G is the greatest of the absolute values of the negative coefficients, then each real root is less than $1 + \sqrt[k]{G/a_0}$.

Example $P(x) = x^5 + 3x^4 - 7x^2 - 40x + 2 = 0$. Here $a_0 = 1$, $G = 40$, and $k = 3$ since we must supply 0 as the coefficient for x^3. Thus $1 + \sqrt[3]{40} \approx 4.42$ is an upper bound for the real roots.

A lower bound to the real roots may be found by applying the criterion to the equation $P(-x)$.

Example $P(-x) = -x^5 + 3x^4 - 7x^2 + 40x + 2 = 0$, which is equivalent to $x^5 - 3x^4 + 7x^2 - 40x - 2 = 0$ since a_0 must be +. Then $a_0 = 1$, $G = 40$, and $k = 1$. Hence $-(1 + 40) = -41$ is a lower bound. Thus all real roots $-41 < r < 4.42$.

One last result is helpful in getting an estimate of how many positive and negative real roots there are.

Descartes Rule The number of positive real roots of a polynomial with real coefficients is either equal to the number of changes in sign v or is less than v by a positive even integer. The number of negative roots of $f(x)$ is either equal to the number of variations of sign of $f(-x)$ or is less than this by a positive even integer.

Example $f(x) = x^4 - 13x^2 + 4x - 2 = 0$ has three changes in sign; therefore, there are either three or one positive roots. $f(-x) = x^4 - 13x^2 - 4x - 2$ has one change in sign. Therefore, there is one negative root.

General Methods for Nonlinear Equations in One Variable

Successive Substitutions Let $f(x) = 0$ be the nonlinear equation to be solved. If this is rewritten as $x = F(x)$, then an iterative scheme can be set up in the form $x_{k+1} = F(x_k)$. To start the iteration an initial guess must be obtained graphically or otherwise. The convergence or divergence of the procedure depends upon the method of writing $x = F(x)$, of which there will usually be several forms. However, if a is a root of $f(x) = 0$, and if $|F'(a)| < 1$, then for any initial approximation sufficiently close to a, the method converges to a. This process is called first order because the error in x_{k+1} is proportional to the first power of the error in x_k for large k.

Example $f(x) = x^3 - x - 1 = 0$. A rough plot shows a real root of approximately 1.3. The equation can be written in the form $x = F(x)$ in several ways such as $x = x^3 - 1$, $x = 1/(x^2 - 1)$, and $x = (1 + x)^{1/3}$. In the first case $F'(x) = 3x^2 = 5.07$ at $x = 1.3$, in the second $F(1.3) = -5.46$, and only in the third case is $F'(1.3) < 1$. Hence only the third iterative process has a chance to converge. This is illustrated in the following table.

Step k	$x = x^3 - 1$	$x = 1/(x^2 - 1)$	$x = (1 + x)^{1/3}$
0	1.3	1.3	1.3
1	1.197	1.4493	1.32
2	0.7150	0.9088	1.3238
3	–0.6345	–5.742	1.3245
4			1.3247

Another way of writing the equation is $x_{k+1} = x_k + \beta f(x_k)$. The choice of β is made such that $|1 + \beta\, df/dx(a)| < 1$. Convergence is guaranteed by the theorem given for simultaneous equations.

Methods of Perturbation Let $f(x) = 0$ be the equation. In general, the iterative relation is

$$x_{k+1} = x_k - [f(x_k)/a_k]$$

where the iteration begins with x_0 as an initial approximation and α_k as some functional.

Newton-Raphson Procedure This variant chooses $\alpha_k = f'(x_k)$ where $f' = df/dx$ and geometrically consists of replacing the graph of $f(x)$ by the tangent line at $x = x_k$ in each successive step. If $f'(x)$ and $f''(x)$ have the same sign throughout an interval $a \le x \le b$ containing the solution, with $f(a)$, $f(b)$ of opposite signs, then the process converges starting from any x_0 in the interval $a \le x \le b$. The process is second order.

Example $f(x) = x - 1 + \dfrac{(0.5)^x - 0.5}{0.3}$

$$f'(x) = 1 - 2.3105[0.5]^x$$

An approximate root (obtained graphically) is 2.

Step k	x_k	$f(x_k)$	$f'(x_k)$
0	2	0.1667	0.4224
1	1.6054	0.0342	0.2407
2	1.4632	0.0055	0.1620

Method of False Position This variant is commenced by finding x_0 and x_1 such that $f(x_0)$, $f(x_1)$ are of opposite signs. Then α_1 = slope of secant line joining $[x_0, f(x_0)]$ and $[x_1, f(x_1)]$ so that

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1)$$

In each following step α_k is the slope of the line joining $[x_k, f(x_k)]$ to the most recently determined point where $f(x_j)$ has the opposite sign from that of $f(x_k)$. This method is of first order. If one uses the most recently determined point (regardless of sign), the method is a secant method.

Method of Wegstein This is a variant of the method of successive substitutions which forces and/or accelerates convergence. The iterative procedure $x_{k+1} = F(x_k)$ is revised by setting $\hat{x}_{k+1} = F(x_k)$ and then taking $x_{k+1} = qx_k + (1 - q)\hat{x}_{k+1}$, where q is a suitably chosen number which may be taken as constant throughout or may be adjusted at each step. Wegstein found that suitable q's are:

Behavior of successive substitution process	Range of optimum q
Oscillatory convergence	$0 < q < ½$
Oscillatory divergence	$½ < q < 1$
Monotonic convergence	$q < 0$
Monotonic divergence	$1 < q$

At each step q may be calculated to give a locally optimum value by setting

$$q = \frac{\hat{x}_{k+1} - \hat{x}_k}{\hat{x}_{k+1} - 2\hat{x}_{k+1} + \hat{x}_{k-1}}$$

The Wegstein method is a secant method applied to $g(x) \equiv x - F(x)$.

Numerical Solution of Simultaneous Nonlinear Equations The techniques illustrated here will be demonstrated for two simultaneous equations $f(x, y) = 0$, $g(x, y) = 0$. They immediately generalize to more than two simultaneous equations.

Method of Successive Substitutions Write a system of equations as

$$\alpha_i = f_i(\alpha), \quad \text{or } \alpha = \mathbf{f}(\alpha)$$

The following theorem guarantees convergence. Let α be the solution to $\alpha_i = f_i(\alpha)$. Assume that given $h > 0$, there exists a number $0 < \mu < 1$ such that

$$\sum_{j=1}^{n} \left| \frac{\partial f_i}{\partial x_j} \right| \le \mu \quad \text{for} \quad |x_i - \alpha_i| < h, \quad i = 1, \ldots, n$$

$$x_i^{k+1} = f_i(x_i^k)$$

Then $x_i^k \to \alpha_i$

as k increases (see Ref. 106).

Newton-Raphson Method To solve the set of equations

$$F_i(x_1, x_2, \ldots, x_n) = 0, \quad \text{or } F_i(\{x_j\}) = 0, \quad \text{or } F_i(\mathbf{x}) = 0$$

one uses a truncated Taylor series to give

$$0 = F_i(\{\mathbf{x}^k\}) + \sum_{j=1}^{n} \left. \frac{\partial F_i}{\partial x_j} \right|_{\mathbf{x}^k} (x_j^{k+1} - x_j^k)$$

Thus one solves iteratively from one point to another.

$$\sum_{j=1}^{n} A_{ij}^k (x_j^{k+1} - x_j^k) = -F_i(\{\mathbf{x}^k\})$$

where $$A_{ij}^k = \left. \frac{\partial F_i}{\partial x_j} \right|_{\mathbf{x}^k}$$

This method requires solution of sets of linear equations until the functions are zero to some tolerance or the changes of the solution between iterations is small enough. Convergence is guaranteed provided the norm of the matrix $\mathbf{A}$ is bounded, $\mathbf{F}(\mathbf{x})$ is bounded for the initial guess, and the second derivative of $\mathbf{F}(\mathbf{x})$ with respect to all variables is bounded. See Refs. 106 and 155.

Example $f(x, y) = 4x^2 + 6x - 4xy + 2y^2 - 3$

$g(x, y) = 2x^2 - 4xy + y^2$

By plotting one of the approximate roots is found to be $x_0 = 0.4$, $y_0 = 0.3$. At this point there results $\partial f/\partial x = 8$, $\partial f/\partial y = -0.4$, $\partial g/\partial x = 0.4$, and $\partial g/\partial y = -1$.

$$8(x^{k+1} - x^k) - 0.4(y^{k+1} - y^k) = +0.26$$

$$0.4(x^{k+1} - x^k) - 1(y^{k+1} - y^k) = -0.07$$

The first few iteration steps are as follows:

Step k	x_k	y_k	$f(x_k, y_k)$	$g(x_k, y_k)$
0	0.4	0.3	−0.26	0.07
1	0.43673	0.24184	0.078	0.0175
2	0.42672	0.25573	−0.0170	−0.007
3	0.42925	0.24943	0.0077	0.0010

Method of Continuity (Homotopy) In the case of n equations in n unknowns, when n is large, determining the approximate solution may involve considerable effort. In such a case the method of continuity is admirably suited for use on digital computers. It consists basically of the introduction of an extra variable into the n equations

$$f_i(x_1, x_2, \ldots, x_n) = 0 \qquad i = 1, \ldots, n \tag{3-72}$$

and replacing them by

$$f_i(x_1, x_2, \ldots, x_n, \lambda) = 0 \qquad i = 1, \ldots, n \tag{3-73}$$

where λ is introduced in such a way that the functions (3-73) depend in a simple way upon λ and reduce to an easily solvable system for $\lambda = 0$ and to the original equations (3-72) for $\lambda = 1$. A system of ordinary differential equations, with independent variable λ, is then constructed by differentiating Eqs. (3-73) with respect to λ. There results

$$\sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j} \frac{dx_j}{d\lambda} + \frac{\partial f_i}{\partial \lambda} = 0 \tag{3-74}$$

where $x_1, \ldots, x_n$ are considered as functions of λ. Equations (3-74) are integrated, with initial conditions obtained from Eqs. (3-73) with $\lambda = 0$, from $\lambda = 0$ to $\lambda = 1$. If the solution can be continued to $\lambda = 1$, the values of $x_1, \ldots, x_n$ for $\lambda = 1$ will be a solution of the original equations. If the integration becomes infinite, the parameter λ must be introduced in a different fashion. Integration of the differential equations (which are usually nonlinear in λ) may be accomplished by using techniques described under "Numerical Solution of Ordinary Differential Equations."

Other Methods Other methods can be found in the literature. See Ref. 66.

INTERPOLATION AND FINITE DIFFERENCES

Linear Interpolation If a function $f(x)$ is approximately linear in a certain range, then the ratio

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} = f[x_0, x_1]$$

is approximately independent of x_0, x_1 in the range. The linear approximation to the function $f(x)$, $x_0 < x < x_1$ then leads to the interpolation formula

$$\begin{aligned} f(x) &\approx f(x_0) + (x - x_0) f[x_0, x_1] \\ &\approx f(x_0) + \frac{x - x_0}{x_1 - x_0} [f(x_1) - f(x_0)] \\ &\approx \frac{1}{x_1 - x_0} [(x_1 - x) f(x_0) - (x_0 - x) f(x_1)] \end{aligned}$$

Example Find cosh 0.83 by linear interpolation given cosh 0.8 and cosh 0.9.

x_i	$f(x_i)$	$x_i - 0.83$
0.8	1.33743	−0.03
0.9	1.43309	+0.07

$$f(0.83) \approx 1/0.10[(0.07)(1.33743) - (-0.03)(1.43309)]$$

$$f(0.83) \approx 1.36613$$

Since the true five-place value is 1.36468, it is seen that here linear interpolation gives three significant figures.

Divided Differences of Higher Order and Higher-Order Interpolation The first-order divided difference $f[x_0, x_1]$ was defined previously. Divided differences of second and higher order are defined iteratively by

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

$$\vdots$$

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_0}$$

and a convenient form for computational purposes is

$$f[x_0, x_1, \ldots, x_k] = \sum_{j=0}^{k}{}' \frac{f(x_j)}{(x_j - x_0)(x_j - x_1) \cdots (x_j - x_k)}$$

for any $k \geq 0$, where the $'$ means that the term $(x_j - x_j)$ is omitted in the denominator. For example,

$$f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

If the accuracy afforded by a linear approximation is inadequate, a generally more accurate result may be based upon the assumption that $f(x)$ may be approximated by a polynomial of degree 2 or higher over certain ranges. This assumption leads to Newton's fundamental interpolation formula with divided differences

$$f(x) \approx f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] + \cdots + (x - x_0)(x - x_1) \cdots (x - x_{n-1}) f[x_0, x_1, \ldots, x_n] + E_n(x)$$

where $$E_n(x) = \text{error} = \frac{1}{(n+1)!} f^{n+1}(\varepsilon)\pi(x)$$

where minimum $(x_0, \ldots, x) < \varepsilon <$ maximum $(x_0, x_1, \ldots, x_n, x)$ and $\pi(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$. In order to use the previous equation most effectively one may first form a divided-difference table. For example, for third-order interpolation the difference table is

where each entry is given by taking the difference between diagonally adjacent entries to the left, divided by the abscissas corresponding to the ordinates intercepted by the diagonals passing through the calculated entry.

Equally Spaced Forward Differences If the ordinates are *equally spaced*, i.e., $x_j - x_{j-1} = \Delta x$ for all j, then the first differences are denoted by $\Delta f(x_0) = f(x_1) - f(x_0)$ or $\Delta y_0 = y_1 - y_0$, where $y = f(x)$. The differences of these first differences, called second differences, are denoted by $\Delta^2 y_0, \Delta^2 y_1, \ldots, \Delta^2 y_n$. Thus

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = y_2 - y_1 - y_1 + y_0 = y_2 - 2y_1 + y_0$$

And in general

$$\Delta^j y_0 = \sum_{n=0}^{j} (-1)^n \binom{j}{n} y_{j-n}$$

where $\binom{j}{n} = \frac{j!}{n!(j-n)!}$ = binomial coefficients.

If the ordinates are equally spaced,

$$x_{n+1} - x_n = \Delta x$$
$$y_n = y(x_n)$$

then the first and second differences are denoted by

$$\Delta y_n = y_{n+1} - y_n$$
$$\Delta^2 y_n = \Delta y_{n+1} - \Delta y_n = y_{n+2} - 2y_{n+1} + y_n$$

A new variable is defined

$$\alpha = \frac{x_\alpha - x_0}{\Delta x}$$

and the finite interpolation formula through the points $y_0, y_1, \ldots, y_n$ is written as follows:

$$y_\alpha = y_0 + \alpha\, \Delta y_0 + \frac{\alpha(\alpha - 1)}{2!} \Delta^2 y_0 + \cdots + \frac{\alpha(\alpha - 1) \cdots (\alpha - n + 1)}{n!} \Delta^n y_0$$

Keeping only the first two terms gives a straight line through (x_0, y_0)–(x_1, y_1); keeping the first three terms gives a quadratic function of position going through those points plus (x_2, y_2). The value $\alpha = 0$ gives $x = x_0$; $\alpha = 1$ gives $x = x_1$, and so on.

Equally Spaced Backward Differences Backward differences are defined by

$$\nabla y_n = y_n - y_{n-1}$$
$$\nabla^2 y_n = \nabla y_n - \nabla y_{n-1} = y_n - 2y_{n-1} + y_{n-2}$$

The interpolation polynomial of order n through the points $y_0, y_{-1}, \ldots, y_{-n}$ is

$$y_\alpha = y_0 + \alpha\, \nabla y_0 + \frac{\alpha(\alpha + 1)}{2!} \nabla^2 y_0 + \cdots + \frac{\alpha(\alpha + 1) \cdots (\alpha + n - 1)}{n!} \nabla^n y_0$$

The value of $\alpha = 0$ gives $x = x_0$; $\alpha = -1$ gives $x = x_{-1}$, and so on. Alternatively, the interpolation polynomial of order n through the points $y_1, y_0, y_{-1}, \ldots, y_{-n}$ is

$$y_\alpha = y_1 + (\alpha - 1)\, \nabla y_1 + \frac{\alpha(\alpha - 1)}{2!} \nabla^2 y_1 + \cdots + \frac{(\alpha - 1)\, \alpha\, (\alpha + 1) \cdots (\alpha + n - 2)}{n!} \nabla^n y_1$$

Now $\alpha = 1$ gives $x = x_1$; $\alpha = 0$ gives $x = x_0$.

Central Differences The central difference denoted by

$$\delta f(x) = f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right)$$

$$\delta^2 f(x) = \delta^{n-1} f\left(x + \frac{h}{2}\right) - \delta^{n-1} f\left(x - \frac{h}{2}\right)$$

is useful for calculating at the interior points of tabulated data.

Lagrange Interpolation Formulas A global polynomial is defined over the entire region of space

$$P_m(x) = \sum_{j=0}^{m} c_j x^j$$

This polynomial is of degree m (highest power is x^m) and order $m + 1$ ($m + 1$ parameters $\{c_j\}$). If we are given a set of $m + 1$ points

$$y_1 = f(x_1), y_2 = f(x_2), \ldots, y_{m+1} = f(x_{m+1})$$

then Lagrange's formula gives a polynomial of degree m that goes through the $m + 1$ points:

$$P_m(x) = \frac{(x - x_2)(x - x_3) \cdots (x - x_{m+1})}{(x_1 - x_2)(x_1 - x_3) \cdots (x_1 - x_{m+1})} y_1 + \frac{(x - x_1)(x - x_3) \cdots (x - x_{m+1})}{(x_2 - x_1)(x_2 - x_3) \cdots (x_2 - x_{m+1})} y_2 + \cdots + \frac{(x - x_1)(x - x_2) \cdots (x - x_{m+1})}{(x_{m+1} - x_1)(x_{m+1} - x_2) \cdots (x_{m+1} - x_m)} y_{m+1}$$

Note that each coefficient of y_j is a polynomial of degree m that vanishes at the points $\{x_j\}$ (except for one value of j) and takes the value of 1.0 at that point:

$$P_m(x_j) = y_j, \quad j = 1, 2, \ldots, m + 1$$

If the function $f(x)$ is known, the error in the approximation is, per Ref. 14,

$$|\text{error}(x)| \leq \frac{|x_{m+1} - x_1|^{m+1}}{(m+2)!} \max_{x_1 \leq x \leq x_{m+1}} |f^{(m+2)}(x)|$$

The evaluation of $P_m(x)$ at a point other than at the defining points can be made with Neville's algorithm (Ref. 231). Let P_1 be the value at x of the unique function passing through the point (x_1, y_1); or $P_1 = y_1$. Let

P_{12} be the value at x of the unique polynomial passing through the points x_1 and x_2. Likewise, $P_{ijk\cdots r}$ is the unique polynomial passing through the points $x_i, x_j, x_k, \ldots x_r$. Then use the table

$$\begin{array}{llllll} x_1 & y_1 = P_1 & & & & \\ & & P_{12} & & & \\ x_2 & y_2 = P_2 & & P_{123} & & \\ & & P_{23} & & P_{1234} & \\ x_3 & y_3 = P_3 & & P_{234} & & \\ & & P_{34} & & & \\ x_4 & y_4 = P_4 & & & & \end{array}$$

These entries are defined using

$$P_{i(i+1)\cdots(i+m)} = \frac{(x - x_{i+m})\,P_{i(i+1)\cdots(i+m-1)} + (x_i - x)\,P_{(i+1)(i+2)\cdots(i+m)}}{x_i - x_{i+m}}$$

For example, consider P_{1234}. The terms on the right-hand side involve P_{123} and P_{234}. The "parents," P_{123} and P_{234}, already agree at points 2 and 3. Here $i = 1$, $m = 3$; thus, the parents agree at $x_{i+1}, \ldots, x_{i+m-1}$ already. The formula makes $P_{i(i+1)\cdots(i+m)}$ agree with the function at the additional points x_{i+m} and x_i. Thus, $P_{i(i+1)\cdots(i+m)}$ agrees with the function at all the points $\{x_i, x_{i+1}, \ldots x_{i+m}\}$.

NUMERICAL DIFFERENTIATION

Numerical differentiation should be avoided whenever possible, particularly when data are empirical and subject to appreciable observation errors. Errors in data can affect numerical derivatives quite strongly; i.e., differentiation is a roughening process. When such a calculation must be made, it is usually desirable first to smooth the data to a certain extent.

Use of Interpolation Formula If the data are given over equidistant values of the independent variable x, an interpolation formula such as the Newton formula (see Refs. 143 and 185) may be used and the resulting formula differentiated analytically. If the independent variable is not at equidistant values, then Lagrange's formulas must be used. By differentiating three- and five-point Lagrange interpolation formulas the following differentiation formulas result for equally spaced tabular points:

Three-Point Formulas Let x_0, x_1, x_2 be the three points.

$$f'(x_0) = \frac{1}{2h}\,[-3f(x_0) + 4f(x_1) - f(x_2)] + \frac{h^2}{3}\,f'''(\varepsilon)$$

$$f'(x_1) = \frac{1}{2h}\,[-f(x_0) + f(x_2)] - \frac{h^2}{6}\,f'''(\varepsilon)$$

$$f'(x_2) = \frac{1}{2h}\,[f(x_0) - 4f(x_1) + 3f(x_2)] + \frac{h^2}{3}\,f'''(\varepsilon)$$

where the last term is an error term $\min_j x_j < \varepsilon < \max_j x_j$.

Smoothing Techniques These techniques involve the approximation of the tabular data by a least-squares fit of the data by using some known functional form, usually a polynomial (for the concept of least squares see "Statistics"). In place of approximating $f(x)$ by a single least-squares polynomial of degree n over the entire range of the tabulation, it is often desirable to replace each tabulated value by the value taken on by a least-squares polynomial of degree n relevant to a subrange of $2M + 1$ points centered, when possible, at the point for which the entry is to be modified. Thus each smoothed value replaces a tabulated value. Let $f_j = f(x_j)$ be the tabular points and y_j = smoothed values.

First-Degree Least Squares with Three Points

$$y_0 = \tfrac{1}{6}[5f_0 + 2f_1 - f_2]$$

$$y_1 = \tfrac{1}{3}[f_0 + f_1 + f_2]$$

$$y_2 = \tfrac{1}{6}[-f_0 + 2f_1 + 5f_2]$$

Second-Degree Least Squares with Five Points For five evenly spaced points x_{-2}, x_{-1}, x_0, x_1, and x_2 (separated by distance h) and their ordinates f_{-2}, f_{-1}, f_0, f_1, and f_2, assume a parabola is fit by least squares. Then the derivative at the center point is

$$f'_0 = 1/10h\,[-2f_{-2} - f_{-1} + f_1 + 2f_2]$$

If derivatives are required at end points, with all points and ordinates to one side, the derivatives are

$$f'_0 = 1/20h\,[-21f_0 + 13f_1 + 17f_2 - 9f_3]$$

$$f'_1 = 1/20h\,[-11f_0 + 3f_1 + 7f_2 + f_3]$$

$$f'_0 = 1/20h\,[21f_0 - 13f_{-1} - 17f_{-2} + 9f_{-1}]$$

$$f'_{-1} = 1/20h\,[11f_0 - 3f_{-1} - 7f_{-2} - f_{-1}]$$

Numerical Derivatives The results given above can be used to obtain numerical derivatives when solving problems on the computer, in particular for the Newton-Raphson method and homotopy methods. Suppose one has a program, subroutine, or other function evaluation device that will calculate f given x. One can estimate the value of the first derivative at x_0 using

$$\left.\frac{df}{dx}\right|_{x_0} \approx \frac{f[x_0(1+\varepsilon)] - f[x_0]}{\varepsilon}$$

(a first-order formula) or

$$\left.\frac{df}{dx}\right|_{x_0} \approx \frac{f[x_0(1+\varepsilon)] - f[x_0(1-\varepsilon)]}{2\varepsilon}$$

(a second-order formula). The value of ε is important; a value of 10^{-6} is typical, but smaller or larger values may be necessary depending on the computer precision and the application. One must also be sure that the value of x_0 is not zero and use a different increment in that case.

NUMERICAL INTEGRATION (QUADRATURE)

A multitude of formulas have been developed to accomplish numerical integration, which consists of computing the value of a definite integral from a set of numerical values of the integrand.

Newton-Cotes Integration Formulas (Equally Spaced Ordinates) for Functions of One Variable The definite integral $\int_a^b f(x)\,dx$ is to be evaluated.

Trapezoidal Rule This formula consists of subdividing the interval $a \le x \le b$ into n subintervals a to $a + h$, $a + h$ to $a + 2h$, . . . and replacing the graph of $f(x)$ by the result of joining the ends of adjacent ordinates by line segments. If $f_j = f(x_j) = f(a + jh), f_0 = f(a), f_n = f(b)$, the integration formula is

$$\int_a^b f(x)\,dx = \frac{h}{2}\,[f_0 + 2f_1 + 2f_2 + \cdots + 2f_{n-1} + f_n] + E_n$$

where $$|E_n| = \frac{nh^3}{12}\,|f''(\varepsilon)| = \frac{(b-a)^3}{12n^2}\,|f''(\varepsilon)| \qquad a < \varepsilon < b$$

This procedure is not of high accuracy. However, if $f''(x)$ is continuous in $a < x < b$, the error goes to zero as $1/n^2$, $n \to \infty$.

Parabolic Rule (Simpson's Rule) This procedure consists of subdividing the interval $a < x < b$ into $n/2$ subintervals, each of length $2h$, where n is an *even* integer. By using the notation as above the integration formula is

$$\int_a^b f(x)\,dx = \frac{h}{3}\,[f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 4f_{n-3} + 2f_{n-2} + 4f_{n-1} + f_n] + E_n$$

where $$|E_n| = \frac{nh^5}{180}\,|f^{(\mathrm{IV})}(\varepsilon)| = \frac{(b-a)^5}{180n^4}\,|f^{(\mathrm{IV})}(\varepsilon)| \qquad a < \varepsilon < b$$

This method approximates $f(x)$ by a parabola on each subinterval. This rule is generally more accurate than the trapezoidal rule. It is the most widely used integration formula.

Gaussian Quadrature Gaussian quadrature provides a highly accurate formula based on irregularly spaced points, but the integral needs to be transformed onto the interval 0 to 1.

$$x = a + (b-a)u, \quad dx = (b-a)du$$

$$\int_a^b f(x)\,dx = (b-a)\int_0^1 f(u)\,du$$

$$\int_0^1 f(u)\,du = \sum_{i=1}^{m} W_i f(u_i)$$

The quadrature is exact when f is a polynomial of degree $2m-1$ in x. Because there are m weights and m Gauss points, we have $2m$ parameters that are chosen to exactly represent a polynomial of degree $2m-1$, which has $2m$ parameters. The Gauss points and weights are given in the table.

Gaussian Quadrature Points and Weights

m	u_i	W_i
2	0.21132 48654	0.50000 00000
	0.78867 51346	0.50000 00000
3	0.11270 16654	0.27777 77778
	0.50000 00000	0.44444 44445
	0.88729 83346	0.27777 77778
4	0.06943 18442	0.17392 74226
	0.33000 94783	0.32607 25774
	0.66999 05218	0.32607 25774
	0.93056 81558	0.17392 74226
5	0.04691 00771	0.11846 34425
	0.23076 53450	0.23931 43353
	0.50000 00000	0.28444 44444
	0.76923 46551	0.23931 43353
	0.95308 99230	0.11846 34425

Example Calculate the value of the following integral.

$$I = \int_0^1 e^{-x} \sin x\,dx \tag{3-75}$$

Using the Gaussian quadrature formulas gives the following values for various values of m. Clearly, three internal points, requiring evaluation of the integrand at only three points, gives excellent results.

m	I
1	0.908185
2	0.910089
3	0.909336367
4	0.909330666
5	0.909330674

Romberg's Method Romberg's method uses extrapolation techniques to improve the answer (Ref. 231). If we let I_1 be the value of the integral obtained using interval size $h = \Delta x$, and I_2 be the value of I obtained when using interval size $h/2$, and I_0 the true value of I, then the error in a method is approximately h^m, or

$$I_1 \approx I_0 + ch^m$$
$$I_2 \approx I_0 + c\left(\frac{h}{2}\right)^m$$

Replacing the ≈ by an equality (an approximation) and solving for c and I_0 gives

$$I_0 = \frac{2^m I_2 - I_1}{2^m - 1}$$

This process can also be used to obtain $I_1, I_2, \ldots$, by halving h each time, and then calculating new estimates from each pair, calling them $J_1, J_2, \ldots$; that is, in the formula above, replace I_0 with J_1. The formulas are reapplied for each pair of J to obtain $K_1, K_2, \ldots$ The process continues until the required tolerance is obtained.

$$\begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \\ & J_1 & J_2 & J_3 \\ & & K_1 & K_2 \\ & & & L_1 \end{array}$$

Romberg's method is most useful for a low-order method (small m) because significant improvement is then possible.

Example Evaluate the same integral (3-75) using the trapezoid rule and then apply the Romberg method. To achieve four-digit accuracy, any result from J_2 through L_1 are suitable, even though the base results (I_1–I_4) are not that close.

$I_1 = 0.967058363$	$I_2 = 0.923704741$	$I_3 = 0.912920511$	$I_4 = 0.910227902$
	$J_1 = 0.909253534$	$J_2 = 0.909325768$	$J_3 = 0.909330366$
		$K_1 = 0.909349846$	$K_2 = 0.909331898$
			$L_1 = 0.909325916$

Singularities When the integrand has singularities, a variety of techniques can be tried. The integral may be divided into one part that can be integrated analytically near the singularity and another part that is integrated numerically. Sometimes a change of argument allows analytical integration. Series expansion might be helpful, too. When the domain is infinite, it is possible to use Gauss-Legendre or Gauss-Hermite quadrature. Also a transformation can be made (Ref. 26). For example, let $u = 1/x$ and then

$$\int_a^b f(x)\,dx = \int_{1/b}^{1/a} \frac{1}{u^2} f\left(\frac{1}{u}\right) du \qquad ab > 0$$

Two-Dimensional Formula Two-dimensional integrals can be calculated by breaking down the integral into one-dimensional integrals.

$$\int_a^b \int_{g_1(x)}^{g_2(x)} f(x, y)\,dx\,dy = \int_a^b G(x)\,dx$$

$$G(x) = \int_{g_1(x)}^{g_2(x)} f(x, y)\,dy$$

Gaussian quadrature can also be used in two dimensions, provided the integration is on a square or can be transformed to one. (Domain transformations might be used to convert the domain to a square.)

$$\int_0^1 \int_0^1 f(x, y)\,dx\,dy = \sum_{i=1}^{mx} W_{xi} \sum_{j=1}^{my} W_{yj} f(x_i, y_j)$$

NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS AS INITIAL VALUE PROBLEMS

A differential equation for a function that depends on only one variable, often time, is called an ordinary differential equation. The general solution to the differential equation includes many possibilities; the boundary or initial conditions are needed to specify which of those are desired. If all conditions are at one point, then the problem is an initial value problem and can be integrated from that point on. If some of the conditions are available at one point and others at another point, then the ordinary differential equations become two-point boundary value problems, which are treated in the next section. Initial value problems as ordinary differential equations arise in control of lumped parameter models, transient models of stirred tank reactors, and in all models where there are no spatial gradients in the unknowns.

A higher-order differential equation

$$y^{(n)} + F(y^{(n-1)}, y^{(n-2)}, \ldots, y', y) = 0$$

with initial conditions

$$G_i(y^{(n-1)}(0), y^{(n-2)}(0), \ldots, y(0), y(0)) = 0, \quad i = 1, \ldots, n$$

can be converted into a set of first-order equations using

$$y_i \equiv y^{(i-1)} = \frac{d^{(i-1)}y}{dt^{(i-1)}} = \frac{d}{dt}\,y^{(i-2)} = \frac{dy_{i-1}}{dt}$$

The higher-order equation can be written as a set of first-order equations.

$$\frac{dy_1}{dt} = y_2$$
$$\frac{dy_2}{dt} = y_3$$

$$\frac{dy_3}{dt} = y_4$$
$$\cdots$$
$$\frac{dy_n}{dt} = -F(y_{n-1}, y_{n-2}, \ldots, y_2, y_1)$$

The initial conditions would have to be specified for variables $y_1(0), \ldots, y_n(0)$, or equivalently $y(0), \ldots, y^{(n-1)}(0)$. The set of equations is then written as

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}, t)$$

All the methods in this section are described for a single equation; the methods apply to multiple equations. See Refs. 106 and 185 for more details.

Euler's method is first-order.

$$y^{n+1} = y^n + \Delta t\, f(y^n)$$

and errors are proportional to Δt. The second-order Adams-Bashforth method is

$$y^{n+1} = y^n + \frac{\Delta t}{2}\,[3f(y^n) - f(y^{n-1})]$$

Errors are proportional to Δt^2, and high-order methods are available. Notice that the higher-order explicit methods require knowing the solution (or the right-hand side) evaluated at times in the past. Since these were calculated to get to the current time, this presents no problem except for starting the problem. Then it may be necessary to use Euler's method with a very small step size for several steps in order to generate starting values at a succession of time points. The error terms, order of the method, function evaluations per step, and stability limitations are listed in Ref. 106. The advantage of the high-order Adams-Bashforth method is that it uses only one function evaluation per step yet achieves high-order accuracy. The disadvantage is the necessity of using another method to start.

Runge-Kutta methods are explicit methods that use several function evaluations for each time step. Runge-Kutta methods are traditionally written for $f(t, y)$. The first-order Runge-Kutta method is Euler's method. A second-order Runge-Kutta method is

$$y^{n+1} = y^n + \frac{\Delta t}{2}\,[f^n + f(t^n + \Delta t, y^n + \Delta t\, f^n)]$$

while the midpoint scheme is also a second-order Runge-Kutta method.

$$y^{n+1} = y^n + \Delta t\, f\left(t^n + \frac{\Delta t}{2}, y^n + \frac{\Delta t}{2} f^n\right)$$

A popular fourth-order Runge-Kutta method is the Runge-Kutta-Feldberg formulas (Ref. 111), which have the property that the method is fourth-order but achieves fifth-order accuracy. The popular integration package RKF45 is based on this method.

$$k_1 = \Delta t\, f(t^n, y^n)$$
$$k_2 = \Delta t\, f\left(t^n + \frac{\Delta t}{4}, y^n + \frac{k_1}{4}\right)$$
$$k_3 = \Delta t\, f\left(t^n + \frac{3}{8}\Delta t, y^n + \frac{3}{32}k_1 + \frac{9}{32}k_2\right)$$
$$k_4 = \Delta t\, f\left(t^n + \frac{12}{13}\Delta t, y^n + \frac{1932}{2197}k_1 - \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3\right)$$
$$k_5 = \Delta t\, f\left(t^n + \Delta t, y^n + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{513}k_3 - \frac{845}{4104}k_4\right)$$
$$k_6 = \Delta t\, f\left(t^n + \frac{\Delta t}{2}, y^n - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2565}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5\right)$$
$$y^{n+1} = y^n + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4104}k_4 - \frac{1}{5}k_5$$
$$z^{n+1} = y^n + \frac{16}{135}k_1 + \frac{6656}{12825}k_3 + \frac{28561}{56430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6$$

The value of $y^{n+1} - z^{n+1}$ is an estimate of the error in y^{n+1} and can be used in step-size control schemes.

Usually one would use a high-order method to achieve high accuracy. The Runge-Kutta-Feldberg method is popular because it is high order and does not require a starting method (as does an Adams-Bashforth method). However, it does require four function evaluations per time step, or four times as many as a fourth-order Adams-Bashforth method. For problems in which the function evaluations are a significant portion of the calculation time, this might be important. Given the speed of present-day computers and the widespread availability of microcomputers (which can be run while you are doing something else, if need be), the efficiency of the methods is most important only for very large problems that are going to be solved many times. For other problems, the most important criterion for choosing a method is probably the time the user spends setting up the problem.

The stability limits for the explicit methods are based on the largest eigenvalue of the linearized system of equations (see Ref. 106). For linear problems, the eigenvalues do not change, so that the stability and oscillation limits must be satisfied for *every* eigenvalue of the matrix **A.** When solving nonlinear problems, the equations are linearized about the solution at the local time, and the analysis applies for small changes in time, after which a new analysis about the new solution must be made. Thus, for nonlinear problems, the eigenvalues keep changing, and the largest stable time step changes, too. The stability limits are:

Euler method, $\lambda\,\Delta t \le 2$
Runge-Kutta, 2nd order, $\lambda\,\Delta t < 2$
Runge-Kutta-Feldberg, $\lambda\,\Delta t < 3.0$

Richardson extrapolation can be used to improve the accuracy of a method. Suppose we step forward one step Δt with a pth-order method. Then redo the problem, this time stepping forward from the same initial point, but in two steps of length $\Delta t/2$, thus ending at the same point. Call the solution of the one-step calculation y_1 and the solution of the two-step calculation y_2. Then an improved solution at the new time is given by

$$y = \frac{2^p y_2 - y_1}{2^p - 1}$$

This gives a good estimate provided Δt is small enough that the method is truly convergent with order p. This process can also be repeated in the same way Romberg's method was used for quadrature.

The error term in the various methods can be used to deduce a step size that will give a user-specified accuracy. Most packages today are based on a user-specified tolerance; the step-size is changed during the calculation to achieve that accuracy. The accuracy itself is not guaranteed, but it improves as the tolerance is decreased.

Implicit Methods By using different interpolation formulas involving y^{n+1}, it is possible to derive implicit integration methods. Implicit methods result in a nonlinear equation to be solved for y^{n+1} so that iterative methods must be used. The backward Euler method is a first-order method.

$$y^{n+1} = y^n + \Delta t\, f(y^{n+1})$$

Errors are proportional to Δt for small Δt. The trapezoid rule is a second-order method.

$$y^{n+1} = y^n + \frac{\Delta t}{2}\,[f(y^n) + f(y^{n+1})]$$

Errors are proportional to Δt^2 for small Δt. When the trapezoid rule is used with the finite difference method for solving partial differential equations, it is called the Crank-Nicolson method. The implicit methods are stable for any step size but do require the solution of a set of nonlinear equations, which must be solved iteratively. The set of equations can be solved using the successive substitution method or Newton-Raphson method. See Ref. 36 for an application to dynamic distillation problems.

The best packages for stiff equations (see below) use Gear's back-

ward difference formulas. The formulas of various orders are, per Refs. 59 and 117,

$$(1)\quad y^{n+1} = y^n + \Delta t\, f(y^{n+1})$$

$$(2)\quad y^{n+1} = \frac{4}{3}y^n - \frac{1}{3}y^{n-1} + \frac{2}{3}\Delta t\, f(y^{n+1})$$

$$(3)\quad y^{n+1} = \frac{18}{11}y^n - \frac{9}{11}y^{n-1} + \frac{2}{11}y^{n-2} + \frac{6}{11}\Delta t\, f(y^{n+1})$$

$$(4)\quad y^{n+1} = \frac{48}{25}y^n - \frac{36}{25}y^{n-1} + \frac{16}{25}y^{n-2} - \frac{3}{25}y^{n-3} + \frac{12}{25}\Delta t\, f(y^{n+1})$$

$$(5)\quad y^{n+1} = \frac{300}{137}y^n - \frac{300}{137}y^{n-1} + \frac{200}{137}y^{n-2} - \frac{75}{137}y^{n-3} + \frac{12}{137}y^{n-4} + \frac{60}{137}\Delta t\, f(y^{n+1})$$

Stiffness The concept of stiffness is described for a system of linear equations.

$$\frac{d\mathbf{y}}{dt} = \mathbf{A}\,\mathbf{y}$$

Let λ_i be the eigenvalues of the matrix **A** (Ref. 267). Then, per Ref. 181, the stiffness ratio is defined as

$$\mathrm{SR} = \frac{\max_i |\mathrm{Re}\,(\lambda_i)|}{\min_i |\mathrm{Re}\,(\lambda_i)|}$$

SR = 20 is not stiff, SR = 10^3 is stiff, and SR = 10^6 is very stiff. If the problem is nonlinear, then the solution is expanded about the current state.

$$\frac{dy_i}{dt} = f_i\,[\mathbf{y}(t^n)] + \sum_{j=1}^{n}\frac{\partial f_i}{\partial y_j}\,[y_j - y_j(t^n)]$$

The question of stiffness then depends on the solution at the current time. Consequently nonlinear problems can be stiff during one time period and not stiff during another. While the chemical engineer may not actually calculate the eigenvalues, it is useful to know that they determine the stability and accuracy of the numerical scheme and the step size used.

Problems are stiff when the time constants for different phenomena have very different magnitudes. Consider flow through a packed bed reactor. The time constants for different phenomena are:

1. Time for device flow-through

$$t_{\mathrm{flow}} = \frac{L}{u} = \frac{\phi AL}{Q}$$

where Q is the volumetric flow rate, A is the cross sectional area, L is the length of the packed bed, and ϕ is the void fraction;

2. Time for reaction

$$t_{r\times n} = \frac{1}{k}$$

where k is a rate constant (time^{-1});

3. Time for diffusion inside the catalyst

$$t_{\mathrm{internal\ diffusion}} = \frac{\varepsilon R^2}{D_e}$$

where ε is the porosity of the catalyst, R is the catalyst radius, and D_e is the effective diffusion coefficient inside the catalyst;

4. Time for heat transfer is

$$t_{\mathrm{internal\ heat\ transfer}} = \frac{R^2}{\alpha} = \frac{\rho_s C_s R^2}{k_e}$$

where ρ_s is the catalyst density, C_s is the catalyst heat capacity per unit mass, k_e is the effective thermal conductivity of the catalyst, and α is the thermal diffusivity. For example, in the model of a catalytic converter for an automobile (Ref. 103), the time constants for internal diffusion was 0.3 seconds; internal heat transfer, 21 seconds; and device flow-through, 0.003 seconds. The device flow-through is so fast that it might as well be instantaneous. The stiffness is approximately 7000. Implicit methods must be used to integrate the equations. Alternatively, a quasistate model can be developed (Ref. 239).

Differential-Algebraic Systems Sometimes models involve ordinary differential equations subject to some algebraic constraints. For example, the equations governing one equilibrium stage (as in a distillation column) are

$$M\frac{dx^n}{dt} = V^{n+1}y^{n+1} - L^n x^n - V^n y^n + L^{n-1}x^{n-1}$$

$$x^{n-1} - x^n = E^n(x^{n-1} - x^{*,n})$$

$$\sum_{i=1}^{N} x_i = 1$$

where x and y are the mole fraction in the liquid and vapor, respectively; L and V are liquid and vapor flow rates, respectively; M is the holdup; and the superscript is the stage number. The efficiency is E, and the concentration in equilibrium with the vapor is x^*. The first equation is an ordinary differential equation for the mass of one component on the stage, while the third equation represents a constraint that the mass fractions add to one. This is a differential-algebraic system of equations.

Differential-algebraic equations can be written in the general notation

$$F\left(t, y, \frac{dy}{dt}\right) = 0$$

To solve the general problem using the backward Euler method, replace the nonlinear differential equation with the nonlinear algebraic equation for one step.

$$F\left(t, y^{n+1}, \frac{y^{n+1} - y^n}{\Delta t}\right) = 0$$

This equation must be solved for y^{n+1}. The Newton-Raphson method can be used, and if convergence is not achieved within a few iterations, the time step can be reduced and the step repeated. In actuality, the higher-order backward-difference Gear methods are used in DASSL (Ref. 224).

Differential-algebraic systems are more complicated than differential systems because the solution may not always be defined. Pontelides et al. (Ref. 226) introduced the term *index* to identify the possible problems. The index is defined as the minimum number of times the equations need to be differentiated with respect to time to convert the system to a set of ordinary differential equations. These higher derivatives may not exist, and the process places limits on which variables can be given initial values. Sometimes the initial values must be constrained by the algebraic equations (Ref. 226). For a differential-algebraic system modeling a distillation tower, Ref. 226 shows that the index depends on the specification of pressure for the column. Byrne and Ponzi (Ref. 58) also list several chemical engineering examples of differential-algebraic systems and solve one involving two-phase flow.

Computer Software Efficient computer packages are available for solving ordinary differential equations as initial value problems. The packages are widely available and good enough that most chemical engineers use them and do not write their own. Here we discuss three of them: RKF45, LSODE, and EPISODE. In each of the packages, the user specifies the differential equation to be solved and a desired error criterion. The package then integrates in time and adjusts the step size to achieve the error criterion within the limitations imposed by stability.

A popular explicit, Runge-Kutta package is RKF45, developed by Forsythe et al. (Ref. 111). The method is based on the Runge-Kutta-Feldberg formulas. Notice there that an estimate of the truncation error at each step is available. Then the step size can be reduced until this estimate is below the user-specified tolerance. The method is thus automatic, and the user is assured of the results. Note, however, that the tolerance is set on the local truncation error, namely from one step to another, whereas the user is usually interested in the global truncation error, or the error after several steps. The global error is generally made smaller by making the tolerance smaller, but the absolute accuracy is not the same as the tolerance. If the problem is stiff, then very

small step sizes are used; the computation becomes very lengthy. The RKF45 code discovers this and returns control to the user with a message indicating the problem is too hard to solve with RKF45.

A popular implicit package is LSODE, a version of Gear's method (Ref. 117) written by Alan Hindmarsh at Lawrence Livermore Laboratory (Ref. 148). In this package, the user specifies the differential equation to be solved and the tolerance desired. Now the method is implicit and therefore stable for any step size. The accuracy may not be acceptable, however, and sets of nonlinear equations must be solved. Thus, in practice the step size is limited but not nearly so much as in the Runge-Kutta methods. In these packages, both the step size and order of the method are adjusted by the package. Suppose we are calculating with a kth order method. The truncation error is determined by the $(k+1)$th order derivative. This is estimated using difference formulas and the values of the right-hand sides at previous times. An estimate is also made for the kth and $(k+2)$th derivative. Then it is possible to estimate the error in a $(k-1)$th order method, a kth order method, and a $(k+1)$th order method. Furthermore, the step size needed to satisfy the tolerance with each of these methods can be determined. Then we can choose the method and step size for the next step that achieves the biggest step, with appropriate adjustments due to the different work required for each order. The package generally starts with a very small step size and a first-order method, the backward Euler method. Then it integrates along, adjusting the order up (and later down) depending on the error estimates. The user is thus assured that the local truncation error meets the tolerance. There is a further difficulty, since the set of nonlinear equations must be solved. Usually a good guess of the solution is available, since the solution is evolving in time and past history can be extrapolated. Thus, the Newton-Raphson method will usually converge. The package protects itself, though, by only doing three iterations. If convergence is not reached within this many iterations, then the step size is reduced and the calculation is redone for that time step. The convergence theorem for the Newton-Raphson method (p. 3-50) indicates that the method will converge if the step size is small enough. Thus the method is guaranteed to work. Further economies are possible. The Jacobian needed in the Newton-Raphson method can be fixed over several time steps. Then, if the iteration does not converge, the Jacobian can be reevaluated at the current time-step. If the iteration still does not converge, then the step-size is reduced and a new Jacobian is evaluated. Also the successive substitution method can be used, which is even faster, except that it may not converge. However, it, too, will converge if the time step is small enough.

Comparisons of the methods and additional details are provided for chemical engineering problems by Refs. 59 and 106. Generally, the Runge-Kutta methods give extremely good accuracy, especially when the step size is kept small for stability reasons. When the computation time is comparable for LSODE and RKF45, the RKF45 package generally gives much more accurate results. The RKF45 package is unsuitable, however, for many chemical reactor problems because they are so stiff. Generally, though, standard packages must have a high-order explicit method (usually a version of Runge-Kutta) and a multistep, implicit method (usually a version of GEAR, EPISODE, or LSODE). The package DASSL (Ref. 224) uses similar principles to solve the differential-algebraic systems.

The software described here is available by electronic mail over the Internet. Sending the message

```
send index to
netlib@ornl.gov
```

will retrieve an index and descriptions of how to obtain the software.

Stability, Bifurcations, Limit Cycles Some aspects of this subject involve the solution of nonlinear equations; other aspects involve the integration of ordinary differential equations; applications include chaos and fractals as well as unusual operation of some chemical engineering equipment. Ref. 176 gives an excellent introduction to the subject and the details needed to apply the methods. Ref. 66 gives more details of the algorithms. A concise survey with some chemical engineering examples is given in Ref. 91. Bifurcation results are closely connected with stability of the steady states, which is essentially a transient phenomenon.

Sensitivity Analysis When solving differential equations, it is frequently necessary to know the solution as well as the sensitivity of the solution to the value of a parameter. Such information is useful when doing parameter estimation (to find the best set of parameters for a model) and for deciding if a parameter needs to be measured accurately. See Ref. 105.

ORDINARY DIFFERENTIAL EQUATIONS-BOUNDARY VALUE PROBLEMS

Diffusion problems in one dimension lead to boundary value problems. The boundary conditions are applied at two different spatial locations: at one side the concentration may be fixed and at the other side the flux may be fixed. Because the conditions are specified at two different locations, the problems are not initial value in character. It is not possible to begin at one position and integrate directly because at least one of the conditions is specified somewhere else and there are not enough conditions to begin the calculation. Thus, methods have been developed especially for boundary value problems.

Shooting Methods The first method is one that utilizes the techniques for initial value problems but allows for an iterative calculation to satisfy all the boundary conditions. Consider the nonlinear boundary value problem

$$\frac{d^2y}{dx^2} = f\left(x, y, \frac{dy}{dx}\right), \quad y(0) = \alpha, \quad y(1) = \beta$$

Convert this second-order equation into two first-order equations along with the boundary conditions written to include a parameter s to represent the unknown value of $v(0) = dy/dx(0)$.

$$\frac{dy}{dx} = v, \quad \frac{dv}{dx} = f(x, y, v), \quad y(0) = \alpha, \quad v(0) = s$$

The parameter s is chosen so that the last boundary condition is satisfied: $y(1) = \beta$. Define the function

$$\chi(s) = y(1, s) - \beta$$

and iterate on s to make $\chi(s) = 0$. Note that the condition at $x = 0$ is satisfied for any s, the differential equation is satisfied by the integration routine, and only the last boundary condition is yet to be satisfied. Both successive substitution and the Newton-Raphson methods can be used. The technique can be used when the boundary conditions are more general and convergence can be proved (see Refs. 106 and 167). Computer software exists: the IMSL program DTPTB uses DVERK, which employs Runge-Kutta integration to integrate the ordinary differential equations (Ref. 55).

Finite Difference Method To apply the finite difference method, we first spread grid points through the domain. Figure 3-49 shows a uniform mesh of n points (nonuniform meshes are possible, too). The unknown, here $c(x)$, at a grid point x_i is assigned the symbol $c_i = c(x_i)$. The finite difference method can be derived easily by using a Taylor expansion of the solution about this point. Expressions for the derivatives are:

$$\left.\frac{dc}{dx}\right|_i = \frac{c_{i+1} - c_i}{\Delta x} - \left.\frac{d^2c}{dx^2}\right|_i \frac{\Delta x}{2} + \cdots, \quad \left.\frac{dc}{dx}\right|_i = \frac{c_i - c_{i-1}}{\Delta x} + \left.\frac{d^2c}{dx^2}\right|_i \frac{\Delta x}{2} + \cdots$$

$$\left.\frac{dc}{dx}\right|_i = \frac{c_{i+1} - c_{i-1}}{2\Delta x} - \left.\frac{d^3c}{dx^3}\right|_i \frac{\Delta x^2}{3!}$$

The truncation error in the first two expressions is proportional to Δx, and the methods are said to be first-order. The truncation error in the third expression is proportional to Δx^2, and the method is said to be second-order. Usually the last equation is used to insure the best accuracy. The finite difference representation of the second derivative is:

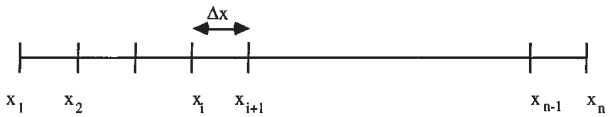

FIG. 3-49 Finite difference mesh; Δx uniform.

$$\left.\frac{d^2c}{dx^2}\right|_i = \frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2} - \left.\frac{d^4c}{dx^4}\right|_i \frac{2\Delta x^2}{4!} + \cdots$$

The truncation error is proportional to Δx^2. To solve a differential equation, it is evaluated at a point i and then these expressions are inserted for the derivatives.

Example Consider the equation for convection, diffusion, and reaction in a tubular reactor.

$$\frac{1}{\text{Pe}}\frac{d^2c}{dx^2} - \frac{dc}{dx} = \text{Da}\, R(c)$$

The finite difference representation is

$$\frac{1}{\text{Pe}}\frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2} - \frac{c_{i+1} - c_{i-1}}{2\Delta x} = \text{Da}\, R(c_i)$$

This equation is written for $i = 2$ to $n - 1$, or the internal points. The equations would then be coupled but would also involve the values of c_1 and c_n, as well. These are determined from the boundary conditions.

If the boundary condition involves a derivative, it is important that the derivatives be evaluated using points that exist. Three possibilities exist:

$$\left.\frac{dc}{dx}\right|_1 = \frac{c_2 - c_1}{\Delta x}$$

$$\left.\frac{dc}{dx}\right|_1 = \frac{-3c_1 + 4c_2 - c_3}{2\Delta x}$$

The third alternative is to add a false point, outside the domain, as $c_0 = c(x = -\Delta x)$.

$$\left.\frac{dc}{dx}\right|_1 = \frac{c_2 - c_0}{2\Delta x}$$

Since this equation introduces a new variable, c_0, another equation is needed and is obtained by writing the finite difference equation for $i = 1$, too.

The sets of equations can be solved using the Newton-Raphson method. The first form of the derivative gives a tridiagonal system of equations, and the standard routines for solving tridiagonal equations suffice. For the other two options, some manipulation is necessary to put them into a tridiagonal form (see Ref. 105).

Frequently, the transport coefficients, such as diffusion coefficient or thermal conductivity, depend on the dependent variable, concentration, or temperature, respectively. Then the differential equation might look like

$$\frac{d}{dx}\left(D(c)\frac{dc}{dx}\right) = 0$$

This could be written as two equations.

$$-\frac{dJ}{dx} = 0 \qquad J = -D(c)\frac{dc}{dx}$$

Because the coefficient depends on c, the equations are more complicated. A finite difference method can be written in terms of the fluxes at the midpoints, $i + 1/2$.

$$-\frac{J_{i+1/2} - J_{i-1/2}}{\Delta x} = 0 \qquad J_{i+1/2} = -D(c_{i+1/2})\frac{c_{i+1} - c_i}{\Delta x}$$

These are combined to give the complete equation.

$$\frac{D(c_{i+1/2})(c_{i+1} - c_i) - D(c_{i-1/2})(c_i - c_{i-1})}{\Delta x^2} = 0$$

This represents a set of nonlinear algebraic equations that can be solved with the Newton-Raphson method. However, in this case, a viable iterative strategy is to evaluate the transport coefficients at the last value and then solve

$$\frac{D(c^k_{i+1/2})(c^{k+1}_{i+1} - c^{k+1}_i) - D(c^k_{i-1/2})(c^{k+1}_i - c^{k+1}_{i-1})}{\Delta x^2} = 0$$

The advantage of this approach is that it is easier to program than a full Newton-Raphson method. If the transport coefficients do not vary radically, then the method converges. If the method does not converge, then it may be necessary to use the full Newton-Raphson method.

There are three common ways to evaluate the transport coefficient at the midpoint. The first one uses the transport coefficient evaluated at the average value of the solutions on either side.

$$D(c_{i+1/2}) \approx D\left[\frac{1}{2}(c_{i+1} + c_i)\right]$$

The truncation error of this approach is Δx^2 (Ref. 106). The second approach uses the average of the transport coefficients on either side.

$$D(c_{i+1/2}) \approx \frac{1}{2}[D(c_{i+1}) + D(c_i)]$$

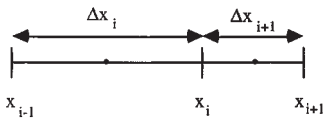

FIG. 3-50 Finite difference grid with variable spacing.

The truncation error of this approach is also Δx^2 (Ref. 106). The third approach uses an "upstream" transport coefficient.

$$D(c_{i+1/2}) \approx D(c_{i+1}), \quad \text{when } D(c_{i+1}) > D(c_i)$$
$$D(c_{i+1/2}) \approx D(c_i), \quad \text{when } D(c_{i+1}) < D(c_i)$$

This approach is used when the transport coefficients vary over several orders of magnitude, and the "upstream" direction is defined as the one in which the transport coefficient is larger. The truncation error of this approach is only Δx (Refs. 106 and 107), but this approach is useful if the numerical solutions show unrealistic oscillations.

If the grid spacing is not uniform, the formulas must be revised. The notation is shown in Fig. 3-50. The finite-difference form of the equations is then

$$-\frac{J_{i+1/2} - J_{i-1/2}}{\frac{1}{2}(\Delta x_i + \Delta x_{i+1})} = 0 \qquad J_{i+1/2} = -D_{i+1/2}\frac{c_{i+1} - c_i}{\Delta x_{i+1}}, \quad J_{i-1/2} = -D_{i-1/2}\frac{c_i - c_{i-1}}{\Delta x_i}$$

If average diffusion coefficients are used, then the finite difference equation is as follows.

$$\frac{1}{\Delta x_{i+1} + \Delta x_i}\left[\frac{1}{\Delta x_{i+1}}(D_{i+1} + D_i)(c_{i+1} - c_i) - \frac{1}{\Delta x_i}(D_i + D_{i-1})(c_i - c_{i-1})\right] = 0$$

Rigorous error bounds are discussed for linear ordinary differential equations solved with the finite difference method by Isaacson and Keller (Ref. 107). Computer software exists to solve two-point boundary value problems. The IMSL routine DVCPR uses the finite difference method with a variable step size (Ref. 247). Finlayson (Ref. 106) gives FDRXN for reaction problems.

Example A reaction diffusion problem is solved with the finite difference method.

$$\frac{d^2c}{dx^2} = \phi^2 c, \quad \frac{dc}{dx}(0) = 0, \quad c(1) = 1$$

The solution is derived for $\phi = 2$. It is solved several times, first with two intervals and three points (at $x = 0$, 0.5, 1), then with four intervals, then with eight intervals. The reason is that when an exact solution is not known, one must use several Δx and see that the solution converges as Δx approaches zero. With two intervals, the equations are as follows. The points are $x_1 = 0$, $x_2 = 0.5$, and $x_3 = 1.0$; and the solution at those points are c_1, c_2, and c_3, respectively. A false boundary is used at $x_0 = -0.5$.

$$\frac{c_0 - c_2}{2\Delta x} = 0, \quad \frac{c_0 - 2c_1 + c_2}{\Delta x^2} - \phi^2 c_1 = 0, \quad \frac{c_1 - 2c_2 + c_3}{\Delta x^2} - \phi^2 c_2 = 0, \quad c_3 = 1$$

The solution is $c_1 = 0.2857$, $c_2 = 0.4286$, and $c_3 = 1.0$. Since the solution is only an approximation and approaches the exact solution only as Δx approaches zero, it is necessary to find out if Δx is small enough to be considered zero. This is done by solving the problem again with more grid points. The value of concentration at $x = 0$ takes the following values for different Δx. These values are extrapolated using the Richardson extrapolation technique to give $c(0) = 0.265826$. Using this value as the best estimate of the exact solution, the errors in the solution are tabulated versus Δx. Clearly the errors go as Δx^2 (decreasing by a factor of 4 when Δx decreases by a factor of 2), thus validating the solution. The exact solution (given below) is 0.265802.

$n - 1$	Δx	$c(0)$
2	0.5	0.285714
4	0.25	0.271043
8	0.125	0.267131

$n - 1$	Δx	Error in $c(0)$
2	0.5	0.01989
4	0.25	0.00521
8	0.125	0.00130

Finite Difference Methods Solved with Spreadsheets A convenient way to solve the finite difference equations for simple

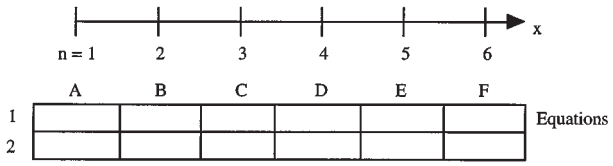

FIG. 3-51 Finite difference method using spreadsheets.

problems is to use a computer spreadsheet. The equations for the problem solved in the example can be cast into the following form

$$c_1 = \frac{2c_2}{2 + \phi^2 \Delta x^2}$$

$$c_i = \frac{c_{i+1} + c_{i-1}}{2 + \phi^2 \Delta x^2}$$

$$c_{n+1} = 1$$

Let us solve the problem using 6 nodes, or 5 intervals. Then the connection between the cell in the spreadsheet and the nodal value is shown in Fig. 3-51. The following equations are placed into the various cells.

```
A1: = 2*B1/(2.+(phi*dx)**2)
B1: = (A1 + C1)/(2.+(phi*dx)**2)
F1: = 1.
```

The equation in cell B1 is copied into cells C1 though E1. Then turn on the iteration scheme in the spreadsheet and watch the solution converge. Whether or not convergence is achieved can depend on how you write the equations, so some experimentation may be necessary. Theorems for convergence of the successive substitution method are useful in this regard.

Orthogonal Collocation The orthogonal collocation method has found widespread application in chemical engineering, particularly for chemical reaction engineering. In the collocation method, the dependent variable is expanded in a series of orthogonal polynomials, and the differential equation is evaluated at certain collocation points. The collocation points are the roots to an orthogonal polynomial, as first used by Lanczos (Refs. 182 and 183). A major improvement was proposed by Villadsen and Stewart (Refs. 288 and 289), who proposed that the entire solution process be done in terms of the solution at the collocation points rather than the coefficients in the expansion. This method is especially useful for reaction-diffusion problems that frequently arise when modeling chemical reactors. It is highly efficient when the solution is smooth, but the finite difference method is preferred when the solution changes steeply in some region of space. See Ref. 105 for comparisons.

Galerkin Finite Element Method In the finite element method, the domain is divided into elements and an expansion is made for the solution on each finite element. In the Galerkin finite element method an additional idea is introduced: the Galerkin method is used to solve the equation. The Galerkin method is explained before the finite element basis set is introduced, using the equations for reaction and diffusion in a porous catalyst pellet.

$$\frac{d^2c}{dx^2} = \phi^2 R(c)$$

$$\frac{dc}{dx}(0) = 0, \quad c(1) = 1$$

The unknown solution is expanded in a series of known functions $\{b_i(x)\}$ with unknown coefficients $\{a_i\}$.

$$c(x) = \sum_{i=1}^{NT} a_i b_i(x)$$

The trial solution is substituted into the differential equation to obtain the residual.

$$\text{Residual} = \sum_{i=1}^{NT} a_i \frac{d^2 b_i}{dx^2} - \phi^2 R\left[\sum_{i=1}^{NT} a_i b_i(x)\right]$$

The residual is then made orthogonal to the set of basis functions.

$$\int_0^1 b_j(x) \left\{\sum_{i=1}^{NT} a_i \frac{d^2 b_i}{dx^2} - \phi^2 R\left[\sum_{i=1}^{NT} a_i b_i(x)\right]\right\} dx = 0 \qquad j = 1, \ldots, NT$$

This is the process that makes the method a Galerkin method. The basis for the orthogonality condition is that a function that is made orthogonal to each member of a complete set is then zero. The residual is being made orthogonal, and if the basis functions are complete and you use infinitely many of them, then the residual is zero. Once the residual is zero, the problem is solved.

This equation is integrated by parts to give the following equation

$$-\sum_{i=1}^{NT} \int_0^1 \frac{db_j}{dx} \frac{db_i}{dx} dx a_i = \phi^2 \int_0^1 b_j(x) R\left[\sum_{i=1}^{NT} a_i b_i(x)\right] dx$$

$$j = 1, \ldots, NT - 1 \quad (3\text{-}76)$$

This equation defines the Galerkin method and a solution that satisfies this equation (for all $j = 1, \ldots, \infty$) is called a weak solution. For an approximate solution, the equation is written once for each member of the trial function, $j = 1, \ldots, NT - 1$, and the boundary condition is applied.

$$\sum_{i=1}^{NT} a_i b_i(1) = c_B$$

The Galerkin finite element method results when the Galerkin method is combined with a finite element trial function. The domain is divided into elements separated by nodes, as in the finite difference method. The solution is approximated by a linear (or sometimes quadratic) function of position within the element. These approximations are substituted into Eq. (3-76) to provide the Galerkin finite element equations. The element integrals are defined as

$$B_{JI}^e = -\frac{1}{\Delta x_e} \int_0^1 \frac{dN_J}{du} \frac{dN_I}{du} du, \quad F_J^e = \phi^2 \Delta x_e \int_0^1 N_J(u) R\left[\sum_{I=1}^{NP} c_I^e N_I(u)\right] du$$

and the entire method can be written in the following compact notation:

$$\sum_e B_{JI}^e c_I^e = \sum_e F_J^e$$

The matrices for various terms are given in the table. This equation can also be written in the form

$$\mathbf{AAc} = \mathbf{f}$$

where the matrix **AA** is sparse; if linear elements are used, the matrix is tridiagonal. Once the solution is found, the solution at any point can be recovered from

$$c^e(u) = c_{I=1}^e (1 - u) + c_{I=2}^e u$$

for linear elements.

Element Matrices for Galerkin Method with Linear Shape Functions

$$N_1 = 1 - u, \; N_2 = u, \; \frac{dN_1}{du} = -1, \; \frac{dN_2}{du} = 1$$

$$\int_0^1 \frac{dN_J}{du} \frac{dN_I}{du} du = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \int_0^1 N_J \frac{dN_I}{du} du = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$\int_0^1 N_J N_I du = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix}, \int_0^1 N_J du = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \int_0^1 N_J u \, du = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{3} \end{pmatrix}$$

Example Solve the specified problem when $\phi = 2$, the rate expression is linear, $R(c) = c$, and the boundary condition is 1.0. The Galerkin finite element method is used with $\Delta x = 0.33333$. The element nodes are at $x = 0$, 0.3333, 0.6667, and 1.0. The solution at $x = 1.0$ is $c_4 = 1.0$. The Galerkin equations for one element are obtained from the table.

$$-\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \phi^2 \Delta x^2 \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix}$$

When these are summed over all elements the result is

$$\begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -1-1 & 1 & 0 \\ 0 & 1 & -1-1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \frac{4}{9} \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & \frac{1}{3}+\frac{1}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{1}{3}+\frac{1}{3} & \frac{1}{6} \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

After rearrangement this is

$$\begin{bmatrix} 31/27 & -25/27 & 0 \\ -25/27 & 62/27 & -25/27 \\ 0 & -25/27 & 62/27 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 25/27 \end{bmatrix}$$

The solution is $c_1 = 0.2560$, $c_2 = 0.3174$, $c_3 = 0.5312$, and $c_4 = 1$. The exact solution is derived using the section entitled "Ordinary Differential Equations: Linear Differential Equations with Constant Coefficients."

$$c = \frac{e^{2x} + e^{-2x}}{e^2 + e^{-2}} = \frac{\cosh(2x)}{\cosh(2)}$$

The values of the exact solution at the same finite element nodes are $c_1 = 0.2658$, $c_2 = 0.3271$, $c_3 = 0.5392$, and $c_4 = 1$, indicating that the three-element finite element solution is accurate within 3 percent. When the exact solution is not known, the problem must be solved several times, each with a different number of elements, so that convergence is seen as the number of elements increases.

Cubic B-Splines Cubic B-splines can also be used to solve differential equations (Refs. 105 and 266).

Adaptive Meshes In many two-point boundary value problems, the difficulty in the problem is the formation of a boundary layer region, or a region in which the solution changes very dramatically. In such cases, it is prudent to use small mesh spacing there, either with the finite difference method or the finite element method. If the region is known *a priori,* small mesh spacings can be assumed at the boundary layer. If the region is not known, though, other techniques must be used. These techniques are known as adaptive mesh techniques. The mesh size is made small where some property of the solution is large. For example, if the truncation error of the method is nth order, then the nth-order derivative of the solution is evaluated and a small mesh is used where it is large. Alternatively, the residual (the differential equation with the numerical solution substituted into it) can be used as a criterion. See Refs. 21 and 107. It is also possible to define the error that is expected from a method one order higher and one order lower. Then a decision about whether to increase or decrease the order of the method can be made, taking into account the relative work of the different orders. This provides a method of adjusting both the mesh spacing (Δx, or sometimes called h) and the degree of polynomial (p). Such methods are called h-p methods.

Singular Problems and Infinite Domains If the solution being sought has a singularity, it may be difficult to find a good numerical solution. Sometimes even the location of the singularity may not be known (Ref. 11). One method of solving such problems is to refine the mesh near the singularity, relying on the better approximation due to a smaller Δx. Another approach is to incorporate the singular trial function into the approximation. Thus, if the solution approaches $f(x)$ as x goes to zero and $f(x)$ becomes infinite, one may define a new variable $u(x) = y(x) - f(x)$ and derive an equation for u. The differential equation is more complicated, but the solution is better near the singularity. See Refs. 39 and 231.

Sometimes the domain is semi-infinite, as in boundary layer flow. The domain can be transformed from the x domain (0–∞) to the η domain (1–0) using the transformation $\eta = \exp(-x)$. Another approach is to use a variable mesh, perhaps with the same transformation. For example, use $\eta = \exp(-\beta x)$ and a constant mesh size in η; the value of β is found experimentally. Still another approach is to solve on a finite mesh in which the last point is far enough away that its location does not influence the solution (Ref. 59). A location that is far enough away must be found by trial and error.

NUMERICAL SOLUTION OF INTEGRAL EQUATIONS

In this subsection is considered a method of solving numerically the Fredholm integral equation of the second kind:

$$u(x) = f(x) + \lambda \int_a^b k(x, t)u(t)\,dt \qquad \text{for } u(x) \tag{3-77}$$

The method discussed arises because a definite integral can be closely approximated by any of several numerical integration formulas (each of which arises by approximating the function by some polynomial over an interval). Thus the definite integral in Eq. (3-77) can be replaced by an integration formula, and Eq. (3-77) may be written

$$u(x) = f(x) + \lambda(b-a)\left[\sum_{i=1}^{n} c_i k(x, t_i)u(t_i)\right] \tag{3-78}$$

where $t_1, \ldots, t_n$ are points of subdivision of the t axis, $a \le t \le b$, and the c's are coefficients whose values depend upon the type of numerical integration formula used. Now Eq. (3-78) must hold for all values of x, $a \le x \le b$; so it must hold for $x = t_1, x = t_2, \ldots, x = t_n$. Substituting for x successively $t_1, t_2, \ldots, t_n$ and setting $u(t_i) = u_i$, $f(t_i) = f_i$, we get n linear algebraic equations for the n unknowns $u_1, \ldots, u_n$. That is,

$$u_i = f_i + (b-a)[c_1 k(t_i, t_1)u_1 + c_2 k(t_i, t_2)u_2 + \cdots + c_n k(t_i, t_n)u_n] \qquad i = 1, 2, \ldots, n$$

These u_j may be solved for by the methods under "Numerical Solution of Linear Equations and Associated Problems" and substituted into Eq. (3-78) to yield an approximate solution for Eq. (3-77).

Example Solve numerically $u(x) = x + \frac{1}{3}\int_0^1 (t + x)u(t)\,dt$. In this example $a = 0$, $b = 1$. Take $n = 3$, $t_1 = 0$, $t_2 = \frac{1}{2}$, $t_3 = 1$. Then Eq. (3-78) takes the form (for which we have used the parabolic rule)

$$u(x) = x + (\tfrac{1}{3})\frac{\frac{1}{2}}{3}[(t_1 + x)u(t_1) + 4(t_2 + x)u(t_2) + (t_3 + x)u(t_3)]$$

$$= x + (1/18)[(t_1 + x)u(t_1) + 4(t_2 + x)u(t_2) + (t_3 + x)u(t_3)]$$

This must hold for all x, $0 \le x \le 1$. Here $t_1 = 0$, $t_2 = \frac{1}{2}$, and $t_3 = 1$. Evaluate at $x = t_i$.

$$u(t_1) = t_1 + \tfrac{1}{18}[2t_1 u(t_1) + 4(t_2 + t_1)u(t_2) + (t_3 + t_1)u(t_3)]$$
$$u(t_2) = t_2 + \tfrac{1}{18}[(t_1 + t_2)u(t_1) + 4(2t_2)u(t_2) + (t_3 + t_2)u(t_3)]$$
$$u(t_3) = t_3 + \tfrac{1}{18}[(t_1 + t_3)u(t_1) + 4(t_2 + t_3)u(t_2) + 2t_3 u(t_3)]$$

By setting in the values of t_1, t_2, t_3 and $u(t_i) = u_i$,

$$18u_1 - 2u_2 - u_3 = 0$$
$$-u_1 + 28u_2 - 3u_3 = 18$$
$$-u_1 - 6u_2 + 16u_3 = 18$$

with the solution $u_1 = 12/71$, $u_2 = 57/71$, $u_3 = 102/71$. Thus

$$u(x) = x + \tfrac{1}{18}[x\,12/71 + 4(\tfrac{1}{2} + x)\,57/71 + (1 + x)\,102/71]$$
$$= 90/71x + 12/71$$

Because of the work involved in solving large systems of simultaneous linear equations it is desirable that only a small number of u's be computed. Thus the gaussian integration formulas are useful because of the economy they offer. See references on numerical solutions of integral equations.

Solutions for Volterra equations are done in a similar fashion, except that the solution can proceed point by point, or in small groups of points depending on the quadrature scheme. See Refs. 105 and 195. There are methods that are analogous to the usual methods for integrating differential equations (Runge-Kutta, predictor-corrector, Adams methods, etc.). Explicit methods are fast and efficient until the time step is very small to meet the stability requirements. Then implicit methods are used, even though sets of simultaneous algebraic equations must be solved. The major part of the calculation is the evaluation of integrals, however, so that the added time to solve the algebraic equations is not excessive. Thus, implicit methods tend to be preferred (Ref. 195). Volterra equations of the first kind are not well posed, and small errors in the solution can have disastrous consequences. The boundary element method uses Green's functions and integral equations to solve differential equations (Refs. 45 and 200).

MONTE CARLO SIMULATIONS

Some physical problems, such as those involving interaction of molecules, are usually formulated as integral equations. Monte Carlo methods are especially well-suited to their solution. This section cannot give a comprehensive treatment of such methods, but their use in

calculating the value of an integral will be illustrated. Suppose we wish to calculate the integral

$$G = \int_{\Omega_0} g(x) f(x)\, dx$$

where the distribution function $f(x)$ satisfies:

$$f(x) \geq 0, \quad \int_{\Omega_0} f(x)\, dx = 1$$

The distribution function $f(x)$ can be taken as constant; for example, $1/\Omega_0$. We choose variables $x_1, x_2, \ldots, x_N$ randomly from $f(x)$ and form the arithmetic mean

$$G_N = \frac{1}{N} \sum_i g(x_i)$$

The quantity G_N is an estimation of G, and the fundamental theorem of Monte Carlo guarantees that the expected value of G_N is G, if G exists (Ref. 161). The error in the calculation is given by

$$\varepsilon = \frac{\sigma_1}{N^{1/2}}$$

where σ_1^2 is calculated from

$$\sigma_1^2 = \int_{\Omega_0} g^2(x) f(x)\, dx - G^2$$

Thus the number of terms needed to achieve a specified accuracy can be calculated once an estimate of σ_1^2 is known.

$$N = \frac{\sigma_1^2}{\varepsilon}$$

Various methods, such as influence sampling, can be used to reduce the number of calculations needed (Ref. 161).

NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

Parabolic Equations in One Dimension By combining the techniques applied to initial value problems and boundary value problems it is possible to easily solve parabolic equations in one dimension. The method is often called the method of lines. It is illustrated here using the finite difference method, but the Galerkin finite element method and the orthogonal collocation method can also be combined with initial value methods in similar ways. The analysis is done by example.

Example Consider the diffusion equation, with boundary and initial conditions.

$$\frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2}$$

$$c(x, 0) = 0$$

$$c(0, t) = 1, \quad c(1, t) = 0$$

We denote by c_i the value of $c(x_i, t)$ at any time. Thus, c_i is a function of time, and differential equations in c_i are ordinary differential equations. By evaluating the diffusion equation at the ith node and replacing the derivative with a finite difference equation, the following working equation is derived for each node i, $i = 2, \ldots, n$ (see Fig. 3-52).

$$\frac{dc_i}{dt} = D \frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2}$$

This can be written in the general form of a set of ordinary differential equations by defining the matrix **AA.**

$$\frac{d\mathbf{c}}{dt} = \mathbf{AAc}$$

This set of ordinary differential equations can be solved using any of the standard methods, and the stability of the integration of these equations is governed by the largest eigenvalue of **AA.** If Euler's method is used for integration, the time step is limited by

$$\Delta t \leq \frac{2}{|\lambda|_{\max}}$$

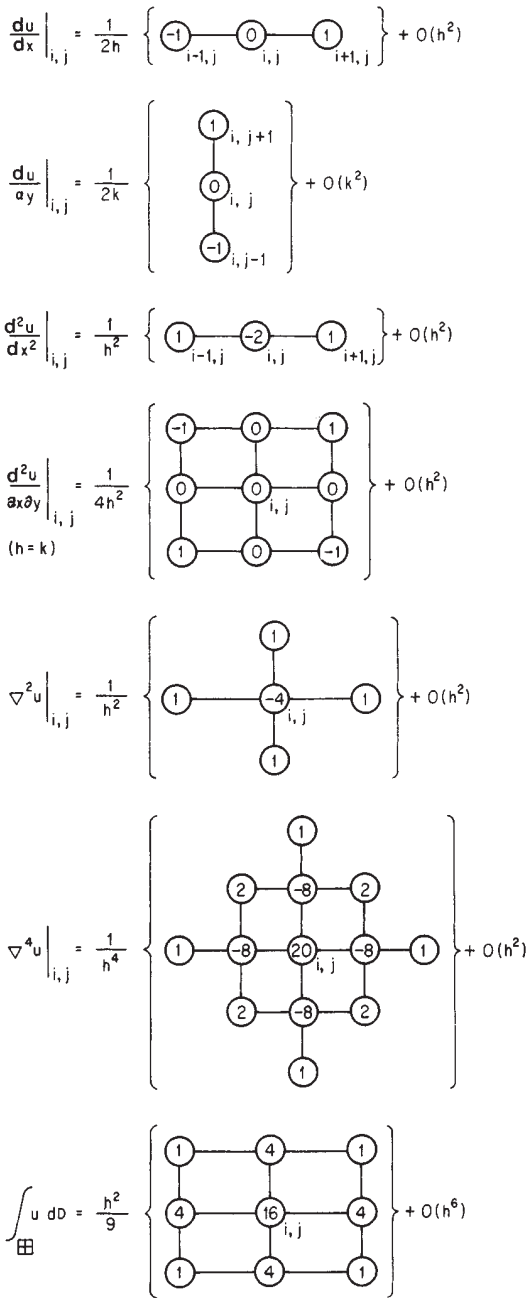

FIG. 3-52 Computational molecules. $h = \Delta x = \Delta y$.

whereas, if the Runge-Kutta-Feldberg method is used, the 2 in the numerator is replaced by 3.0. The largest eigenvalue of **AA** is bounded by Gerschgorin's Theorem (Ref. 155, p. 135).

$$|\lambda|_{\max} \leq \max_{2<j<n} \sum_{i=2}^{n} |\mathbf{AA}_{ji}| = \frac{4D}{\Delta x^2}$$

This gives the well-known stability limit

$$\Delta t \frac{D}{\Delta x^2} \leq \frac{1}{2}$$

The smallest eigenvalue is independent of Δx (it is $D\pi^2/L^2$) so that the ratio of largest to smallest eigenvalue is proportional to $1/\Delta x^2$. Thus, the problem becomes stiff as Δx approaches zero (Ref. 106).

Another way to study the stability of explicit equations is to use the positivity theorem. For Euler's method, the equations can be written in the form

$$\frac{c_i^{n+1} - c_i^n}{\Delta t} = D\,\frac{c_{i+1}^n - 2c_i^n + c_{i-1}^n}{\Delta x^2}$$

where $c_i^n = c(x_i, t^n)$. Then the new value is given by

$$c_i^{n+1} = \frac{D\Delta t}{\Delta x^2}\,c_{i+1}^n + \left(1 - 2\,\frac{D\Delta t}{\Delta x^2}\right)c_i^n + \frac{D\Delta t}{\Delta x^2}\,c_{i-1}^n$$

Theorem. If $c_i^{n+1} = Ac_{i+1}^n + Bc_i^n + Cc_{i-1}^n$ and A, B, and C are positive and $A + B + C \le 1$, then the scheme is stable and the errors die out. Here the theorem requires

$$\left(1 - 2\,\frac{D\Delta t}{\Delta x^2}\right) > 0$$

which gives the same stability condition (Ref. 106).

Implicit methods can also be used. Write a finite difference form for the time derivative and average the right-hand sides, evaluated at the old and new time.

$$\frac{c_i^{n+1} - c_i^n}{\Delta t} = D(1-\theta)\,\frac{c_{i+1}^n - 2c_i^n + c_{i-1}^n}{\Delta x^2} + D\theta\,\frac{c_{i+1}^{n+1} - 2c_i^{n+1} + c_{i-1}^{n+1}}{\Delta x^2}$$

Now the equations are of the form

$$-\frac{D\Delta t\theta}{\Delta x^2}\,c_{i+1}^{n+1} + \left[\,1 + 2\,\frac{D\Delta t\theta}{\Delta x^2}\right]c_i^{n+1} - \frac{D\Delta t\theta}{\Delta x^2}\,c_{i-1}^{n+1} = c_i^n + \frac{D\Delta t(1-\theta)}{\Delta x^2}\,(c_{i+1}^n - 2c_i^n + c_{i-1}^n)$$

and require solving a set of simultaneous equations, which have a tridiagonal structure. Using $\theta = 0$ gives the Euler method (as above), $\theta = 0.5$ gives the Crank-Nicolson method, and $\theta = 1$ gives the backward Euler method. The Crank-Nicolson method is also the same as applying the trapezoid rule to do the integration. The stability limit is given by

$$\frac{D\Delta t}{\Delta x^2} \le \frac{0.5}{1 - 2\theta}$$

If the Δt satisfies the following equation, then the solution will not oscillate from node to node (a numerical artifact). See Ref. 106.

$$\frac{D\Delta t}{\Delta x^2} \le \frac{0.25}{1 - \theta}$$

Other methods can be used in space, such as the finite element method, the orthogonal collocation method, or the method of orthogonal collocation on finite elements (see Ref. 106). Spectral methods employ Chebyshev polynomials and the Fast Fourier Transform and are quite useful for hyperbolic or parabolic problems on rectangular domains (Ref. 125).

Packages exist that use various discretizations in the spatial direction and an integration routine in the time variable. PDECOL uses B-splines for the spatial direction and various GEAR methods in time (Ref. 247). PDEPACK and DSS (Ref. 247) use finite differences in the spatial direction and GEARB in time (Ref. 66). REACOL (Ref. 106) uses orthogonal collocation in the radial direction and LSODE in the axial direction, while REACFD uses finite difference in the radial direction; both codes are restricted to modeling chemical reactors.

Elliptic Equations Elliptic equations can be solved with both finite difference and finite element methods. One-dimensional elliptic problems are two-point boundary value problems. Two- and three-dimensional elliptic problems are often solved with iterative methods when the finite difference method is used and direct methods when the finite element method is used. So there are two aspects to consider: how the equations are discretized to form sets of algebraic equations and how the algebraic equations are then solved.

The prototype elliptic problem is steady-state heat conduction or diffusion,

$$k\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) = Q$$

possibly with a heat generation term per unit volume, Q. The boundary conditions taken here are $T = f(x, y)$ on the boundary (S) with f a known function. Illustrations are given for constant thermal conductivity k while Q is a known function of position. The finite difference formulation is given using the following nomenclature:

$$T_{i,j} = T(i\Delta x, j\Delta y)$$

The finite difference formulation is then (see Fig. 3-52)

$$\frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{\Delta x^2} + \frac{T_{i,j+1} - 2T_{i,j} + T_{i,j-1}}{\Delta y^2} = Q_{i,j} \qquad (3\text{-}79)$$

$$T_{i,j} = f(x_i, y_j) \text{ on } S$$

If the boundary is parallel to a coordinate axis any derivative is evaluated as in the section on boundary value problems, using either a one-sided, centered difference or a false boundary. If the boundary is more irregular and not parallel to a coordinate line then more complicated expressions are needed and the finite element method may be the better method.

Equation (3-79) is rewritten in the form

$$2\left(1 + \frac{\Delta x^2}{\Delta y^2}\right)T_{i,j} = T_{i+1,j} + T_{i-1,j} + \frac{\Delta x^2}{\Delta y^2}\,(T_{i,j+1} + T_{i,j-1}) - \Delta x^2\,\frac{Q_{i,j}}{k}$$

The relaxation method solves this equation iteratively.

$$2\left(1 + \frac{\Delta x^2}{\Delta y^2}\right)T_{i,j}^* = T_{i+1,j}^s + T_{i-1,j}^{s+1} + \frac{\Delta x^2}{\Delta y^2}\,(T_{i,j+1}^s + T_{i,j-1}^{s+1}) - \Delta x^2\,\frac{Q_{i,j}}{k}$$

$$T_{i,j}^{s+1} = T_{i,j}^s + \beta(T_{i,j}^* - T_{i,j}^s)$$

If $\beta = 1$, this is the Gauss-Seidel method. If $\beta > 1$, it is overrelaxation; if $\beta < 1$ it is underrelaxation. The value of β may be chosen empirically, $0 < \beta < 2$, but it can be selected theoretically for simple problems like this (Refs. 106 and 221). In particular, these equations can be programmed in a spreadsheet and solved using the iteration feature, provided the boundaries are all rectangular.

The alternating direction method can be used for elliptic problems by using sequences of iteration parameters (Refs. 106 and 221). The method is well suited to transient problems as well.

These are the classical iterative techniques. Recently preconditioned conjugate gradient methods have been developed (see Ref. 100). In these methods, a series of matrix multiplications are done iteration by iteration; and the steps lend themselves to the efficiency available in parallel computers. In the multigrid method, the problem is solved on several grids, each more refined than the previous one. As one iterates between the solutions on the different grids, one converges to the solution of the algebraic equations. See Juncu and Mihail (Ref. 68) for a chemical engineering application.

The Galerkin finite element method (FEM) is useful for solving elliptic problems and is particularly effective when the domain or geometry is irregular. As an example, cover the domain with triangles and define a trial function on each triangle. The trial function takes the value 1.0 at one corner and 0.0 at the other corners and is linear in between. See Fig. 3-53. These trial functions on each triangle are pieced together to give a trial function on the whole domain. General treatments of the finite element method are available (see references). The steps in the solution method are similar to those described for boundary value problems, except now the problems are much bigger so that the numerical analysis must be done very carefully to be efficient. Most engineers, though, just use a finite element program without generating it. There are three major caveats that must be addressed. The first one is that the solution is dependent on the mesh laid down, and the only way to assess the accuracy of the solution is to solve the problem with a more refined mesh. The second concern is that the solution obeys the shape of the trial function inside

the element. Thus, if linear functions are used on triangles, a three-dimensional view of the solution, plotting the solution versus x and y, consists of a series of triangular planes joined together at the edges, as in a geodesic dome. The third caveat is that the Galerkin finite element method is applied to both the differential equations and the boundary conditions. Computer programs are usually quite general and may allow the user to specify boundary conditions that are not realistic. Also, natural boundary conditions are satisfied if no other boundary condition (ones involving derivatives) is set at a node. Thus, the user of finite element codes must be very clear what boundary conditions and differential equations are built into the computer code. When the problem is nonlinear, the Newton-Raphson method is used to iterate from an initial guess. Nonlinear problems lead to complicated integrals to evaluate, and they are usually evaluated using Gaussian quadrature.

One nice feature of the finite element method is the use of natural boundary conditions. It may be possible to solve the problem on a domain that is shorter than needed to reach some limiting condition (such as at an outflow boundary). The externally applied flux is still applied at the shorter domain, and the solution *inside* the truncated domain is still valid. Examples are given in Refs. 67 and 107. The effect of this is to allow solutions in domains that are smaller, thus saving computation time and permitting the solution in semi-infinite domains.

A general purpose package for general two-dimensional domains and rectangular three-dimensional rectangular domains is ELLPACK (Ref. 247). This package allows choice of a variety of methods: finite difference, Hermite collocation, spline Galerkin, collocation, as well as others. Comparisons of the various methods are available (Ref. 154). The program FISHPAK solves the Helmholtz equation in multiple dimensions when the domain is separable (since fast methods like FFT are used). See Ref. 247.

Hyperbolic Equations The most common situation yielding hyperbolic equations involves unsteady phenomena with convection. Two typical equations are the convective diffusive equation

$$\frac{\partial c}{\partial t} + u\,\frac{\partial c}{\partial x} = D\,\frac{\partial^2 c}{\partial x^2}$$

and the chromatography equation (Ref. 245)

$$\phi\,\frac{\partial c}{\partial t} + \phi u\,\frac{\partial c}{\partial x} + (1-\phi)\,\frac{df}{dc}\,\frac{\partial c}{\partial t} = 0$$

where ϕ is the void fraction and $f(c)$ gives the equilibrium relation between the concentration in the fluid phase and the concentration in the solid phase. If the diffusion coefficient is zero, the convective diffusion equation is hyperbolic. If D is small, the phenomenon may be essentially hyperbolic, even though the equations are parabolic. Thus the numerical methods for hyperbolic equations may be useful even for parabolic equations.

Equations for several methods are given here, as taken from the book by Finlayson (Ref. 107). If the convective term is treated with a centered difference expression, the solution exhibits oscillations from node to node, and these only go away if a very fine grid is used. The simplest way to avoid the oscillations with a hyperbolic equation is to use upstream derivatives. If the flow is from left to right, this would give

$$\frac{dc_i}{dt} + u\,\frac{c_i - c_{i-1}}{\Delta x} = D\,\frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2}$$

$$\frac{d}{dt}\,[\,\phi c_i + (1-\phi) f(c_i)] + \phi u_i\,\frac{c_i - c_{i-1}}{\Delta x} = 0$$

(See Ref. 227 for the reason the equation is written in this form.)

The effect of using upstream derivatives is to add artificial or numerical diffusion to the model. This can be ascertained by rearranging the finite difference form of the convective diffusion equation

$$\frac{dc_i}{dt} + u\,\frac{c_{i+1} - c_{i-1}}{2\Delta x} = \left(D + \frac{u\Delta x}{2}\right)\frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2}$$

Thus the diffusion coefficient has been changed from

$$D \text{ to } D + \frac{u\Delta x}{2}$$

Another method often used for hyperbolic equations is the MacCormack method. This method has two steps, and it is written here for the convective diffusion equation.

$$c_i^{*n+1} = c_i^n - \frac{u\Delta t}{\Delta x}\,(c_{i+1}^n - c_i^n) + \frac{D\Delta t}{\Delta x^2}\,(c_{i+1}^n - 2c_i^n + c_{i-1}^n)$$

$$c_i^{n+1} = \frac{1}{2}\,(c_i^n + c_i^{*n+1}) - \frac{u\Delta t}{2\Delta x}\,(c_i^{*n+1} - c_{i-1}^{*n+1})$$

$$+ \frac{D\Delta t}{2\Delta x^2}\,(c_{i+1}^{*n+1} - 2c_i^{*n+1} + c_{i-1}^{*n+1})$$

The concentration profile is steeper for the MacCormack method than for the upstream derivatives, but oscillations can still be present. The flux-corrected transport method can be added to the MacCormack method. A solution is obtained both with the upstream algorithm and the MacCormack method and then they are combined to add just enough diffusion to eliminate the oscillations without smoothing the solution too much. The algorithm is complicated and lengthy but well worth the effort (Refs. 37, 107, and 270).

Stability conditions can be constructed in terms of $\text{Co} = u\Delta t/\Delta x$ and $r = D\Delta t/\Delta x^2$ by using Fourier analysis (Ref. 107). All the methods require

$$\text{Co} = \frac{u\Delta t}{\Delta x} \le 1$$

where Co is the Courant number. How much Co should be less than one depends on the method and on $r = D\Delta t/\Delta x^2$. For example, the upstream method requires $\text{Co} \le 1 - 2r$. The MacCormack method depends less on r and is stable for most Co as long as $r \le 0.5$. Each of these methods is trying to avoid oscillations that would disappear if the mesh were fine enough. For the steady convective diffusion equation, these oscillations *do not* occur provided

$$\frac{u\Delta x}{2D} \le 1$$

For large velocity u, the Δx must be small to meet this condition. An alternative is to use a small Δx in regions where the solution changes drastically. Since these regions change in time, it is necessary that the elements or grid points move. The criteria to move the grid points can be quite complicated, and typical methods are reviewed in Ref. 107. Similar considerations apply to the nonlinear chromatography problem (Ref. 227). See especially Ref. 192.

Parabolic Equations in Two or Three Dimensions Computations become much more lengthy when there are two or more spatial dimensions. For example, we may have the unsteady heat conduction equation

$$\rho C_p\,\frac{\partial T}{\partial t} = k\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) - Q$$

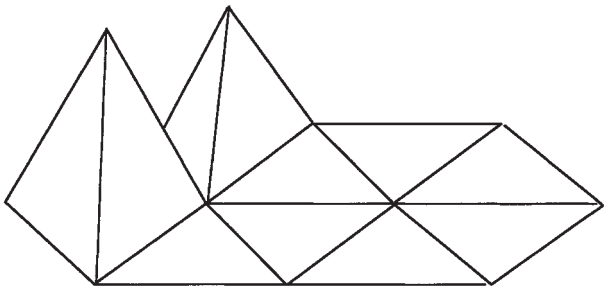

FIG. 3-53 Trial functions for Galerkin finite element method: linear polynomial on triangle.

In the finite difference method an explicit technique would evaluate the right-hand side at the nth time level.

$$\rho C_p \frac{T_{i,j}^{n+1} - T_{i,j}^n}{\Delta t} = \frac{k}{\Delta x^2}(T_{i+1,j}^n - 2T_{i,j}^n + T_{i-1,j}^n) + \frac{k}{\Delta y^2}(T_{i,j+1}^n - 2T_{i,j}^n + T_{i,j-1}^n) - Q$$

When $Q = 0$ and $\Delta x = \Delta y$, the time step limit can be found using the positivity rule.

$$\Delta t \le \frac{\Delta x^2 \rho C_p}{4k} \qquad \text{or} \qquad \frac{\Delta x^2}{4D}$$

These time steps are smaller than for one-dimensional problems. For three dimensions, the limit is

$$\Delta t \le \frac{\Delta x^2}{6D}$$

To avoid such small time steps, which become smaller as Δx decreases, an implicit method could be used. This leads to large, sparse matrices rather than convenient tridiagonal matrices. These can be solved, but the alternating direction method is also useful (Ref. 221). This reduces a problem on an $n \times n$ grid to a series of $2n$ one-dimensional problems on an n grid.

SPLINE FUNCTIONS

Splines are functions that match given values at the points $x_1, \ldots, x_{NT}$ and have continuous derivatives up to some order at the knots, or the points $x_2, \ldots, x_{NT-1}$. Cubic splines are most common; see Ref. 38. The function is represented by a cubic polynomial within each interval (x_i, x_{i+1}) and has continuous first and second derivatives at the knots. Two more conditions can be specified arbitrarily. These are usually the second derivatives at the two end points, which are commonly taken as zero; this gives the natural cubic splines.

Take $y_i = y(x_i)$ at each of the points x_i, and let $\Delta x_i = x_{i+1} - x_i$. Then, in the interval (x_i, x_{i+1}), the function is represented as a cubic polynomial.

$$C_i(x) = a_{0i} + a_{1i}x + a_{2i}x^2 + a_{3i}x^3$$

The interpolating function takes on specified values at the knots and has continuous first and second derivatives at the knots. Within the ith interval, the function is

$$C_i(x) = C_i(x_i) + C_i'(x_i)(x - x_i) + C_i''(x_i)\frac{(x - x_i)}{2} + [C_i''(x_{i+1}) - C_i''(x_i)]\frac{(x - x_i)^3}{6\Delta x_i}$$

where $C_i(x_i) = y_i$. The second derivative $C_i''(x_i) = y_i''$ is found by solving the following tridiagonal system of equations:

$$y_{i-1}''\Delta x_{i-1} + y_i''2(\Delta x_{i-1} + \Delta x_i) + y_{i+1}''\Delta x_i = 6\left(\frac{y_i - y_{i-1}}{\Delta x_{i-1}} - \frac{y_{i+1} - y_i}{\Delta x_i}\right)$$

Since the continuity conditions apply only for $i = 2, \ldots, NT - 1$, we have only $NT - 2$ conditions for the NT values of y_i''. Two additional conditions are needed, and these are usually taken as the value of the second derivative at each end of the domain, y_1'', y_{NT}''. If these values are zero, we get the natural cubic splines; they can also be set to achieve some other purpose, such as making the first derivative match some desired condition at the two ends. With these values taken as zero in the natural cubic spline, we have a $NT - 2$ system of tridiagonal equations, which is easily solved. Once the second derivatives are known at each of the knots, the first derivatives are given by

$$y_i' = \frac{y_{i+1} - y_i}{\Delta x_i} - y_i''\frac{\Delta x_i}{3} - y_{i+1}''\frac{\Delta x_i}{6}$$

The function itself is then known within each element.

FAST FOURIER TRANSFORM (REF. 231)

Suppose a signal $y(t)$ is sampled at equal intervals

$$y_n = y(n\Delta), \qquad n = \ldots, -2, -1, 0, 1, 2, \ldots$$

Δ = sampling rate (e.g., number of samples per second)

The Fourier transform and inverse transform are

$$Y(\omega) = \int_{-\infty}^{\infty} y(t)e^{i\omega t}\,dt$$

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Y(\omega)e^{-i\omega t}\,dt$$

The Nyquist critical frequency or critical angular frequency is

$$f_c = \frac{1}{2\Delta}, \qquad \omega_c = \frac{\pi}{\Delta}$$

If a function $y(t)$ is bandwidth-limited to frequencies smaller than f_c, such as

$$Y(\omega) = 0 \qquad \text{for } \omega > \omega_c$$

then the function is completely determined by its samples y_n. Thus, the entire information content of a signal can be recorded by sampling at a rate $\Delta^{-1} = 2f_c$. If the function is *not* bandwidth-limited, then aliasing occurs. Once a sample rate Δ is chosen, information corresponding to frequencies greater than f_c is simply aliased into that range. The way to detect this in a Fourier transform is to see if the transform approaches zero at $\pm f_c$; if not, aliasing has occurred, and a higher sampling rate is needed.

Next, suppose we have N samples, where N is even

$$y_k = y(t_k) \qquad t_k = k\Delta \qquad k = 0,1,2, \ldots, N-1$$

and the sample rate is Δ. With only N values $\{y_k\}$, it is not possible to determine the complete Fourier transform $Y(\omega)$. We calculate the value $Y(\omega_n)$ at the discrete points

$$\omega_n = \frac{2\pi n}{N\Delta}, \quad n = -\frac{N}{2}, \ldots, 0, \ldots, \frac{N}{2}$$

$$Y_n = \sum_{k=0}^{N-1} y_k e^{2\pi i k n/N}$$

$$Y(\omega_n) = \Delta Y_n$$

The discrete inverse Fourier transform is

$$y_k = \frac{1}{N}\sum_{n=0}^{N-1} Y_n e^{-2\pi i k n/N}$$

The fast Fourier transform (FFT) is used to calculate the Fourier transform as well as the inverse Fourier transform. A discrete Fourier transform of length N can be written as the sum of two discrete Fourier transforms, each of length $N/2$.

$$Y_k = Y_k^e + W^k Y_k^o$$

Here Y_k is the kth component of the Fourier transform of y, and Y_k^e is the kth component of the Fourier transform of the even components of $\{y_j\}$ and is of length $N/2$. Similarly, Y_k^o is the kth component of the Fourier transform of the odd components of $\{y_j\}$ and is of length $N/2$. W is a constant, which is taken to the kth power.

$$W = e^{2\pi i/N}$$

Since Y_k has N components, while Y_k^e and Y_k^o have $N/2$ components, Y_k^e and Y_k^o are repeated once to give N components in the calculation of Y_k. This decomposition can be used recursively. Thus, Y_k^e is split into even and odd terms of length $N/4$.

$$Y_k^e = Y_k^{ee} + W^k Y_k^{eo}$$

$$Y_k^o = Y_k^{oe} + W^k Y_k^{oo}$$

This process is continued until there is only one component. For this reason, the number N is taken as a power of 2. The vector $\{y_j\}$ is filled with zeroes, if need be, to make $N = 2^p$ for some p. For the computer program, see Ref. 26. The standard Fourier transform takes N^2 operations to calculation, whereas the fast Fourier transform takes only $N \log_2 N$. For large N, the difference is significant; at $N = 100$ it is a factor of 15, but for $N = 1000$ it is a factor of 100.

The discrete Fourier transform can also be used for differentiating a function, and this is used in the spectral method for solving differential equations. Suppose we have a grid of equidistant points

$$x_n = n\Delta x, \quad n = 0, 1, 2, \ldots, 2N-1, \qquad \Delta x = \frac{L}{2N}$$

The solution is known at each of these grid points $\{Y(x_n)\}$. First the Fourier transform is taken.

$$y_k = \frac{1}{2N} \sum_{n=0}^{2N-1} Y(x_n) e^{-2ik\pi x_n/L}$$

The inverse transformation is

$$Y(x) = \frac{1}{L} \sum_{k=-N}^{N} y_k e^{2ik\pi x/L}$$

Differentiate this to get

$$\frac{dY}{dx} = \frac{1}{L} \sum_{k=-N}^{N} y_k \frac{2\pi ik}{L} e^{2ik\pi x/L}$$

Thus at the grid points

$$\left.\frac{dY}{dx}\right|_n = \frac{1}{L} \sum_{k=-N}^{N} y_k \frac{2\pi ik}{L} e^{2ik\pi x_n/L}$$

The process works as follows. From the solution at all grid points the Fourier transform is obtained using FFT, $\{y_k\}$. Then this is multiplied by $2\pi ik/L$ to obtain the Fourier transform of the derivative.

$$y_k = y_k \frac{2\pi ik}{L}$$

Then the inverse Fourier transform is taken using FFT, giving the value of the derivative at each of the grid points.

$$\left.\frac{dY}{dx}\right|_n = \frac{1}{L} \sum_{k=-N}^{N} y_k e^{2ik\pi x_n/L}$$

OPTIMIZATION

INTRODUCTION

Optimization should be viewed as a tool to aid in decision making. Its purpose is to aid in the selection of better values for the decisions that can be made by a person in solving a problem. To formulate an optimization problem, one must resolve three issues. First, one must have a representation of the artifact that can be used to determine how the artifact performs in response to the decisions one makes. This representation may be a mathematical model or the artifact itself. Second, one must have a way to evaluate the performance—an objective function—which is used to compare alternative solutions. Third, one must have a method to search for the improvement. This section concentrates on the third issue, the methods one might use. The first two items are difficult ones, but discussing them at length is outside the scope of this section.

Example optimization problems are: (1) determining the optimal thickness of pipe insulation; (2) finding the best equipment sizes and operating schedules for the design of a new batch process to make a given slate of products; (3) choosing the best set of operating conditions for a set of experiments to determine the constants in a kinetic model for a given reaction; (4) finding the amounts of a given set of ingredients one should use for making a carbon rod to be used as an electrode in an arc welder.

For the first problem, one will usually write a mathematical model of how insulation of varying thicknesses restricts the loss of heat from a pipe. Evaluation requires that one develop a cost model for the insulation (a capital cost in dollars) and the heat that is lost (an operating cost in dollars/year). Some method is required to permit these two costs to be compared, such as a present worth analysis. Finally, if the model is simple enough, the method one can use is to set the derivative of the evaluation function to zero with respect to wall thickness to find candidate points for its optimal thickness. For the second problem, selecting a best operating schedule involves discrete decisions, which will generally require models that have integer variables.

It may not be possible to develop a mathematical model for the fourth problem if not enough is known to characterize the performance of a rod versus the amounts of the various ingredients used in its manufacture. The rods may have to be manufactured and judged by ranking the rods relative to each other, perhaps based partially or totally on opinions. Pattern search methods have been devised to attack problems in this class.

In this section assume a mathematical model is possible for the problem to be solved. The model may be encoded in a subroutine and be known only implicitly, or the equations may be known explicitly. A general form for such an optimization problem is

$$\min F = F(z), \text{ such that } h(z) = 0 \text{ and } g(z) \le 0$$

where F represents a specified objective function that is to be minimized. Functions h and g represent equality and inequality constraints that must be satisfied at the final problem solution.

Variables z are used to model such things as flows, mole fractions, physical properties, temperatures, and sizes. The objective function F is generally assumed to be a scalar function, one which represents such things as cost, net present value, safety, or flexibility. Sometimes several objective functions are specified (e.g., minimizing cost while maximizing reliability); these are commonly combined into one function, or else one is selected for the optimization while the others are specified as constraints. Equations $h(z) = 0$ are typically algebraic equations, linear or nonlinear, when modeling steady-state processes, or algebraic coupled with ordinary and/or partial differential equations when optimizing time-varying processes. Inequalities $g(z) \le 0$ put limits on the values variables can take, such as a minimum and maximum temperature, or they restrict one pressure to be greater than another.

An important issue is how to solve large problems that occur in distributed systems. The optimization of distributed systems is discussed in Refs. 52, 120, 244, and 285. For further reading on optimization, readers are directed to Refs. 120 and 244 as well as introductory texts on optimization applied to chemical engineering (Refs. 99 and 225). The material in this section is part of a more advanced treatment (Ref. 295).

Packages There are a number of packages available for optimization, some of which are listed here.

1. *Frameworks*

• *GAMS.* This framework is commercially available. It provides a uniform language to access several different optimization packages, many of them listed below. It will convert the model as expressed in "GAMS" into the form needed to run the package chosen.

• *AMPL.* This framework is by Fourier and coworkers (Ref. 113) at Northwestern University. It is well suited for constructing complex models.

• *ASCEND.* This framework is by Westerberg and coworkers (Ref. 295) at Carnegie-Mellon University. It features an object-oriented modeling language and is well suited for constructing complex models.

2. *Algebraic optimization with equality and inequality constraints*

• *SQP.* A package by Biegler at Carnegie-Mellon University.

• *MINOS5.4.* A package available from Stanford Research Institute (affiliated with Stanford University). This package is the state of the art for mildly nonlinear programming problems.

• *GRG.* A package from Lasdon at the University of Texas, Dept. of Management Science.

3. *Linear programming.* Most current commercial codes for lin-

ear programming extend the Simplex algorithm, and they can typically handle problems with up to 15,000 constraints.

- *MPSX.* From IBM
- *SCICONIC.* From the company of that name
- *MINOS5.4*
- *Cplex.* A package by R. Bixby at Rice University and Cplx, Inc.

CONDITIONS FOR OPTIMALITY

Local Minimum Point for Unconstrained Problems Consider the following unconstrained optimization problem:

$$\min_u \{F(u) \mid u \in \mathbf{R}^r\}$$

If F is continuous and has continuous first and second derivatives, it is necessary that F is stationary with respect to all variations in the independent variables u at a point $\hat{u}$, which is proposed as a minimum to F; that is,

$$\frac{\partial F}{\partial u_i} = 0, \quad i = 1, 2, \ldots, r \qquad \text{or} \qquad \nabla_u F = 0 \quad \text{at } u = \hat{u} \tag{3-80}$$

These are only necessary conditions, as point $\hat{u}$ may be a minimum, maximum, or saddle point.

Sufficient conditions are that any local move away from the optimal point $\hat{u}$ gives rise to an increase in the objective function. Expand F in a Taylor series locally around the candidate point $\hat{u}$ up to second-order terms:

$$F(u) = F(\hat{u}) + \nabla_u F^T\big|_{\hat{u}} (u - \hat{u}) + \frac{1}{2} (u - \hat{u})^T \nabla^2_{uu} F\big|_{\hat{u}} (u - \hat{u}) + \cdots$$

If $\hat{u}$ satisfies necessary conditions [Eq. (3-80)], the second term disappears in this last line. Sufficient conditions for the point to be a local minimum are that the matrix of second partial derivatives $\nabla^2_{uu} F$ is positive definite. This matrix is symmetric, so all of its eigenvalues are real; to be positive definite, they must all be greater than zero.

Constrained Derivatives—Equality Constrained Problems Consider minimizing the objective function F written in terms of n variables z and subject to m equality constraints $h(z) = 0$, or

$$\min_z \{F(z) \mid h(z) = 0, z \in \mathbf{R}^n, h{:}\mathbf{R}^n \to \mathbf{R}^m\} \tag{3-81}$$

The point $\hat{z}$ is tested to see if it could be a minimum point. It is necessary that F be stationary for all infinitesimal moves for z that satisfy the equality constraints. Linearize the m equality constraints around $\hat{z}$, getting

$$h(\hat{z} + \Delta z) = h(\hat{z}) + \nabla_z h^T\big|_{\hat{z}} \Delta z \tag{3-82}$$

where $\Delta z = z - \hat{z}$. There are m constraints here, so m of the variables are dependent, leaving $r = n - m$ independent variables. Partition the variables Δz into a set of m dependent variables Δx and $r = n - m$ independent variables Δu. Equation (3-82), rearranged and then rewritten in terms of these variables, becomes

$$\Delta h = \nabla_x h^T\big|_{\hat{z}} \Delta x + \nabla_u h^T\big|_{\hat{z}} \Delta u = 0$$

This enables the solution for Δx. Linearize the objective function $F(z)$ in terms of the partitioned variables

$$\Delta F = \nabla_x F^T\big|_{\hat{z}} \Delta x + \nabla_u F^T\big|_{\hat{z}} \Delta u$$

and substitute for Δx.

$$\Delta F = \{\nabla_x F^T - \nabla_u F^T [\nabla_x h^T]^{-1} \nabla_u h^T\}_{\hat{z}} \Delta u$$

$$= \left\{\frac{dF}{du}\right\}^T_{\Delta h = 0} \Delta u = \sum_{i=1}^{r} \left\{\frac{dF}{du_i}\right\}_{\Delta h = 0} \Delta u_i$$

There is one term for each Δu_i in the row vector which is in the curly braces {}. These terms are called **constrained derivatives,** which tells how the object function changes when the independent variables u_i are changed while keeping the constraints satisfied (by varying the dependent variables x_i).

Necessary conditions for optimality are that these constrained derivatives are zero; that is,

$$\left\{\frac{dF}{du_i}\right\}_{\Delta h = 0} = 0, \quad i = 1, 2, \ldots, r$$

Equality Constrained Problems—Lagrange Multipliers Form a scalar function, called the Lagrange function, by adding each of the equality constraints multiplied by an arbitrary multiplier to the objective function.

$$L(x, u, \lambda) = F(x, u) + \sum_{i=1}^{m} \lambda_i h_i(x, u) = F(x, u) + \lambda^T h(x, u)$$

At any point where the functions $h(z)$ are zero, the Lagrange function equals the objective function.

Next differentiate L with respect to variables x, u, and λ.

$$\nabla_x L^T\big|_{\hat{z}} = \nabla_x F^T\big|_{\hat{z}} + \lambda^T \nabla h_x^T\big|_{\hat{z}} = 0^T \tag{3-83}$$

$$\nabla_u L^T\big|_{\hat{z}} = \nabla_u F^T\big|_{\hat{z}} + \lambda^T \nabla h_u^T\big|_{\hat{z}} = 0^T \tag{3-84}$$

$$\nabla_\lambda L^T\big|_{\hat{z}} = h^T(x, u) = 0^T$$

Solve Eq. (3-83) for the Lagrange multipliers

$$\lambda^T = -\nabla_x F^T [\nabla h_x^T]^{-1} \tag{3-85}$$

and then eliminate these multipliers from Eq. (3-84).

$$\nabla_u L^T = \nabla_u F^T - \nabla_x F^T [\nabla h_x^T] \nabla h_u^T = 0^T$$

$\nabla_u L$ is equal to the constrained derivatives for the problem, which should be zero at the solution to the problem. Also, these stationarity conditions very neatly provide the necessary conditions for optimality of an equality-constrained problem.

Lagrange multipliers are often referred to as shadow prices, adjoint variables, or dual variables, depending on the context. Suppose the variables are at an optimum point for the problem. Perturb the variables such that only constraint h_i changes. We can write

$$\Delta L = \Delta F + \lambda_i \Delta h_i = 0$$

which is zero because, as just shown, the Lagrange function is at a stationary point at the optimum. Solving for the change in the objective function:

$$\Delta F = -\lambda_i \Delta h_i$$

The multiplier tells how the optimal value of the objective function changes for this small change in the value of a constraint while holding all the other constraints at zero. It is for this reason that they are often called shadow prices.

Equality- and Inequality-Constrained Problems—Kuhn-Tucker Multipliers Next a point is tested to see if it is an optimum one when there are inequality constraints. The problem is

$$\min_z \{F(z) \mid h(z) = 0, g(z) \le 0, z \in \mathbf{R}^n, F{:}\mathbf{R}^n \to \mathbf{R}^1, h{:}\mathbf{R}^n \to \mathbf{R}^m, g{:}\mathbf{R}^n \to \mathbf{R}^p\}$$

The Lagrange function here is similar to that used above.

$$L(z, \lambda, \mu) \equiv F(z) + \lambda^T h(z) + \mu^T g(z)$$

Each of the inequality constraints $g_i(z)$ multiplied by what is called a Kuhn-Tucker multiplier μ_i is added to form the Lagrange function. The necessary conditions for optimality, called the Karush-Kuhn-Tucker conditions for inequality-constrained optimization problems, are

$$\begin{aligned}
&\nabla_z L\big|_{\hat{z}} = \nabla_z F\big|_{\hat{z}} + \nabla_z h\big|_{\hat{z}} \lambda + \nabla_z g\big|_{\hat{z}} \mu = 0 \\
&\nabla_\lambda L = h(z) = 0 \\
&g(z) \le 0 \\
&\mu_i g_i(z) = 0, \quad i = 1, 2, \ldots, p \\
&\mu_i \ge 0, \quad i = 1, 2, \ldots, p
\end{aligned} \tag{3-86}$$

Conditions in Eq. (3-86), called complementary slackness conditions, state that either the constraint $g_i(z) = 0$ and/or its corresponding multiplier μ_i is zero. If constraint $g_i(z)$ is zero, it is behaving like an equality constraint, and its multiplier μ_i is exactly the same as a Lagrange multiplier for an equality constraint. If the constraint is

away from zero, it is not a part of the problem and should not affect it. Setting its multiplier to zero removes it from the problem.

As the goal is to minimize the objective function, releasing the constraint into the feasible region must not decrease the objective function. Using the shadow price argument above, it is evident that the multiplier must be nonnegative (Ref. 177).

Sufficiency conditions to assure that a Kuhn-Tucker point is a local minimum point require one to prove that the objective function will increase for any feasible move away from such a point. To carry out such a test, one has to generate the matrix of second derivatives of the Lagrange function with respect to all the variables z evaluated at $\hat{z}$. The test is seldom done, as it requires too much work.

STRATEGIES OF OPTIMIZATION

The theory just covered tells if a candidate point is or is not the optimum point, but how is the candidate point found? The simplest strategy is to place a grid of points throughout the feasible space, evaluating the objective function at every grid point. If the grid is fine enough, then the point yielding the highest value for the objective function can be selected as the optimum. Twenty variables gridded over only ten points would take place over 10^{20} points in our grid, and, at one nanosecond per evaluation, it would take in excess of four thousand years to carry out these evaluations.

Most strategies limit themselves to finding a local minimum point in the vicinity of the starting point for the search. Such a strategy will find the global optimum only if the problem has a single minimum point or a set of "connected" minimum points. A "convex" problem has only a global optimum.

Pattern Search Suppose the optimization problem is to find the right mix of a given set of ingredients and the proper baking temperature and time to make the best cake possible. A panel of judges can be formed to judge the cakes; assume they are only asked to rank the cakes and that they can do that task in a consistent manner. Our approach will be to bake several cakes and ask the judges to rank them. For this type of problem, pattern-search methods can be used to find the better conditions for manufacturing the product. We shall only describe the ideas behind this approach. Details on implementing it can be found in Ref. 284.

The complex method is one such pattern search method (see Fig. 3-54). First, form a "complex" of at least $r + 1$ ($r = 2$ and 4 points are used in Fig. 3-54) different points at which to bake the cakes by picking a range of suitable values for the r independent variables for the baking process. Bake the cakes and then ask the judges to identify the worst cake.

For each independent variable, form the average value at which it was run in the complex. Draw a line from the coordinates of the worst cake through the average point—called the centroid—and continue on that line a distance that is twice that between these two points. This point will be the next test point. First decide if it is feasible. If so, bake the cake and discover if it leads to a cake that is better than the worst cake from the last set of cakes. If it is not feasible or it is not better, then return half the distance toward the average values from the last test and try again. If it is better, toss out the worst point of the last test and replace it with this new one. Again, ask the judges to find the worst cake. Continue as above until the cakes are all the same quality in the most recent test. It might pay to restart at this point, stopping finally if the restart leads to no improvement. The method takes large steps if the steps are being successful in improving the recipe. It collapses onto a set of points quite close to each other otherwise. The method works reasonably well, but it requires one to bake lots of cakes.

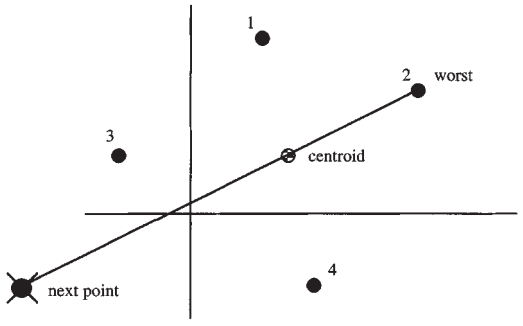

FIG. 3-54 Complex method, a pattern search optimization method.

The following strategies are all examples of Generalized Reduced Gradient (GRG) methods.

Optimization of Unconstrained Objective Assume the objective function F is a function of independent variables u_i, $i = 1 \cdots r$. A computer program, given the values for the independent variables, can calculate F and its derivatives with respect to each u_i. Assume that F is well approximated as an as-yet-unknown quadratic function in u.

$$F \approx a + b^T u + \frac{1}{2} u^T Q u$$

where a is a scalar; b, a vector; and Q, an $r \times r$ symmetric positive definite matrix. The gradient of the approximate function is

$$\nabla_u F = b + Qu$$

Setting the gradient to zero allows an estimate for its minimum.

$$u = -Q^{-1} b \tag{3-87}$$

Initially, Q and b are not known and the calculation proceeds as follows: b contains r unknown coefficients and Q another $r(r + 1)/2$. To estimate b and Q, the computer code is used repeatedly, getting r equations each time—namely

$$\begin{aligned} (\nabla_u F)(1) &= b + Qu(1) \\ (\nabla_u F)(2) &= b + Qu(2) \\ &\cdots \\ (\nabla_u F)(t) &= b + Qu(t) \end{aligned} \tag{3-88}$$

As soon as there are as many independent equations as there are unknown coefficients, these *linear* equations are solved for b and Q. A proper choice of the points $u(i)$ guarantees getting independent equations to solve here.

Given b and Q, Eq. (3-87) provides a new estimate for u as a candidate minimum point. The subroutine is used again to obtain the gradient of F at this point. If the gradient is essentially zero, the calculations stop, since a point has been found that satisfies the necessary conditions for optimality. If not, the equations are written in the form of Eq. (3-88) for this new point, adding them to the set while removing the oldest set of equations. The new set of equations for b and Q are solved, and the calculations continue until a minimum point is found. If removal of the oldest equations from the set in Eq. (3-88) leads to a singular set of equations, then different equations have to be selected for removal. Alternatively, all the older equations can be kept, with the new ones added to the top of the list. Pivoting can be done by proceeding down the list until a nonsingular set of equations is found. Then the older equations are used only if necessary. Also, since only one set of equations is being replaced, clever methods are available to find the solution to the equations with much less work than is required to solve the set of equations the first time (Refs. 89 and 259).

Quadratic Fit for the Equality Constrained Case Next consider solving a problem of the form of Eq. (3-82). For each iteration k:

1. Enter with values provided for variables $u(k)$.
2. Given values for $u(k)$, solve equations $h(x, u) = 0$ for $x(k)$. These will be m equations in m unknowns. If the equations are nonlinear, solving can be done using a variant of the Newton-Raphson method.
3. Use Eq. (3-85) to solve for the Lagrange multipliers $\lambda(k)$. If the Newton-Raphson method (or any or several variants to it) is used to solve the equations, the jacobian matrix $\nabla_x^T h|_{z(k)}$ and its LU factors are already known so solving Eq. (3-85) requires very little effort.
4. Substitute $\lambda(k)$ into Eq. (3-84), which in general will not be zero. The gradient $\nabla_u L(k)$ computed will be the constrained derivatives of F with respect to the independent variables $u(k)$.
5. Return.

The calculations begin with given values for the independent variables u and exit with the (constrained) derivatives of the objective function with respect to them. Use the routine described above for the unconstrained problem where a succession of quadratic fits is used to move toward the optimal point for an unconstrained problem. This approach is a form of the generalized reduced gradient (GRG) approach to optimizing, one of the better ways to carry out optimization numerically.

Inequality Constrained Problems To solve inequality constrained problems, a strategy is needed that can decide which of the inequality constraints should be treated as equalities. Once that question is decided, a GRG type of approach can be used to solve the resulting equality constrained problem. Solving can be split into two phases: phase 1, where the goal is to find a point that is feasible with respect to the inequality constraints; and phase 2, where one seeks the optimum while maintaining feasibility. Phase 1 is often accomplished by ignoring the objective function and using instead

$$F = \sum_{i=1}^{p} \begin{Bmatrix} g_i^2(z) \text{ if } g_i(z) > 0) \\ 0 \text{ otherwise} \end{Bmatrix}$$

until all the inequality constraints are satisfied.

Then at each point, check which of the inequality constraints are active, or exactly equal to zero. These can be placed into the active set and treated as equalities. The remaining can be put aside to be used only for testing. A step can then be proposed using the GRG algorithm. If it does not cause one to violate any of the inactive inequality constraints, the step is taken. Otherwise one can add the closest inactive inequality constraint to the active set. Finding the closet inactive equality will almost certainly require a line search in the direction proposed by the GRG algorithm.

When one comes to a stationary point, one has to test the active inequality constraints at that point to see if they should remain active. This test is done by examining the sign (they should be nonnegative if they are to remain active) of their respective Kuhn-Tucker multipliers. If any should be released, it has to be done carefully as the release of a constraint changes the multipliers for all the constraints. One can find oneself cycling through the testing to decide whether to release the constraints. A correct approach is to add slack variables s to the problem to convert the inequality constraints to equalities and then require the slack variables to remain positive. The multipliers associated with the inequalities $s \geq 0$ all behave independently, and their sign tells one directly to keep or release the constraints. In other words, simultaneously release all the slack variables that have multipliers strictly less than zero. If released, *the slack variables must be treated as a part of the set of independent variables* until one is well away from the associated constraints for this approach to work.

Successive Quadratic Programming (SQP) The above approach to finding the optimum is called a feasible path method, as it attempts at all times to remain feasible with respect to the equality and inequality constraints as it moves to the optimum. A quite different method exists called the Successive Quadratic Programming (SQP) method, which only requires one be feasible at the final solution. Tests that compare the GRG and SQP methods generally favor the SQP method so it has the reputation of being one of the best methods known for nonlinear optimization for the type of problems considered here.

Assume certain inequality constraints will be active at the final solution. The necessary conditions for optimality are

$$\nabla_z L(z, \mu, \lambda) = \nabla F + \nabla g_A \mu + \nabla h \lambda = 0, \quad g_A(z) = 0, \quad h(z) = 0$$

Then one can apply Newton's method to the necessary conditions for optimality, which are a set of simultaneous (non)linear equations. The Newton equations one would write are

$$\begin{bmatrix} \nabla_{zz} L[z(i), u(i), \lambda(i)] & \nabla g_A[z(i)] & \nabla h[z(i)] \\ \nabla g_A[z(i)]^T & 0 & 0 \\ \nabla h[z(i)]^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta z(i) \\ \Delta \mu(i) \\ \Delta \lambda(i) \end{bmatrix} = - \begin{bmatrix} \nabla_z L[z(i), \mu(i), \lambda(i)] \\ g_A[z(i)] \\ h[z(i)] \end{bmatrix}$$

A sufficient condition for a unique Newton direction is that the matrix of constraint derivatives is of full rank (linear independence of constraints) and the Hessian matrix of the Lagrange function $[\nabla_{zz} L(z, \mu, \lambda)]$ *projected into the space of the linearized constraints* is positive definite. The linearized system actually represents the solution of the following quadratic programming problem:

$$\min_{\Delta z} \nabla F[z(i)]^T \Delta z + \frac{1}{2} \Delta z^T \nabla_{zz} L[z(i), \mu(i), \lambda(i)] \Delta z$$

subject to

$$g_A[z(i)] + \nabla g_A[z(i)]^T \Delta z = 0 \qquad \text{and} \qquad h[z(i)] + \nabla h[z(i)]^T \Delta z = 0$$

Reformulating the necessary conditions as a linear quadratic program has an interesting side effect. We can simply add linearizations of the inactive inequalities to the problem and let the active set be selected by the algorithm used to solve the linear quadratic program.

Problems with calculating second derivatives as well as maintaining positive definiteness of the Hessian matrix can be avoided by approximating this matrix by $B(i)$ using a quasi-Newton formula such as BFGS (Refs. 50, 84, 109, 110, 122, and 259). One maintains positive definiteness by skipping the update if it causes the matrix to lose this property. Here gradients of the Lagrange function are used to calculate the update formula (Refs. 136 and 228). The resulting quadratic program, which generates the search direction at each iteration i, becomes:

$$\min_{\Delta z} \nabla F[z(i)]^T \Delta z + \frac{1}{2} \Delta z^T B(i) \Delta z$$

subject to

$$g[z(i)] + \nabla g[z(i)]^T \Delta z \leq 0$$

$$h[z(i)] + \nabla h[z(i)]^T \Delta z = 0$$

This linear quadratic program will have a unique solution if $B(i)$ is kept positive definite. Efficient solution methods exist for solving it (Refs. 119 and 123).

Finally, to ensure convergence of this algorithm from poor starting points, a step size α is chosen along the search direction so that the point at the next iteration ($z^{i+1} = z^i + \alpha d$) is closer to the solution of the NLP (Refs. 65, 136, and 254).

These problems get very large as the Lagrange function involves all the variables in the problem. If one has a problem with 5000 variables z and the problem has only 10 degrees of freedom (i.e., the partitioning will select 4990 variables x and only 10 variables u), one is still faced with maintaining a matrix B that is 5000×5000. See Westerberg (Ref. 40) for references to this case.

Interior Point Algorithms for Linear Programming Problems There has been considerable excitement in the popular press about so-called interior point algorithms (Ref. 23) for solving extremely large linear programming problems. Computational demands for these algorithms grow less rapidly than for the Simplex algorithm, with a break-even point being a few thousand constraints. A key idea for an interior method is that one heads across the feasible region to locate the solution rather than around its edges as one does for the Simplex algorithm. This move is found by computing the direction of steepest descent for the objective function with respect to changing the slack variables. Variables u are computed in terms of the slack variables by using the inequality constraints. The direction of steepest descent is a function of the scaling of the variables used for the problem. See Refs. 6, 124, 199, and 295.

Linear Programming The combined term *linear programming* is given to any method for finding where a given linear function of several variables takes on an extreme value, and what that value is, when the variables are nonnegative and are constrained by linear equalities or inequalities. A very general problem consists of maximizing $f = \sum_{j=1}^{n} c_j z_j$ subject to the constraints $z_j \geq 0$ ($j = 1, 2, \ldots, n$) and $\sum_{j=1}^{n} a_{ij} z_j \leq b_i$ ($i = 1, 2, \ldots, m$). With S the set of all points whose coordinates z_j satisfy all the constraints, we must ask three questions: (1) Are the constraints *consistent?* If not, S is empty and there is no solution. (2) If S is not empty, does the function f become *unbounded* on S? If so, the problem has no solution. If not, then there is a point P of S that is optimal in the sense that if Q is any point of S then $f(Q) \leq f(P)$. (3) How can we find P?

The simplex algorithm, in a sense, prepares the problem before cal-

culation in such a way that favorable answers to these questions are tentatively assumed for the given problem and can be guaranteed for the prepared problem. The calculations then reveal whether or not those assumptions are justified for the given problem. The simplex algorithm terminates automatically, yielding full information on the given problem and so-called dual problem. The dual of the general problem of linear programming is to minimize $d(\mu_1, \ldots, \mu_m) = \sum_{i=1}^{m} \mu_i b_i$ subject to $\mu_i \geq 0$ $(i = 1, 2, \ldots, m)$ and $\sum_{i=1}^{m} \mu_i a_{ij} \geq c_j (j = 1, 2, \ldots, n)$. Let A be the matrix $[a_{ij}]$, $c = [c_j]$, $U = [\mu_i]$ be row vectors, and $B = [b_i]^T$, $Z = [z_j]^T$ be column vectors. In matrix form the original (primal) problem is to maximize $f(Z) = CZ$ subject to $Z \geq 0$, $AZ \leq B$. The dual is to minimize $d(U) = UB$ subject to $U \geq 0$, $UA \geq C$.

Example Maximize $3z_1 + 4z_2$ subject to the constraints $z_1 \geq 0$, $z_2 \geq 0$, $2z_1 + 4z_2 \leq 8$, and $4z_1 + 3z_2 \leq 10$. The dual problem is to minimize $8\mu_1 + 10\mu_2$ subject to the constraints $\mu_1 \geq 0$, $\mu_2 \geq 0$, $2\mu_1 + 4\mu_2 \geq 3$, and $5\mu_1 + 3\mu_2 \geq 4$.

Simplex Method

1. *Original problem.* Let the column vector $[z_j]^T = z$ $(j = 1, 2, \ldots, n)$ and the row vector $[c_j] = c$. To maximize $f(z) = \sum_{j=1}^{n} c_j z_j = c^T z$ subject to the n constraints $z_j \geq 0$ $(j = 1, \ldots, n)$ and m further constraints h_i: $\sum_{j=1}^{n} a_{ij} z_j \circ_i b_i$ $(i = 1, 2, \ldots, m)$ where $\circ_i$ can be $\geq$ = or $\leq$. If any $b_i \leq 0$, multiply h_i by -1; thus we may assume $b_i \geq 0$. We suppose the m constraints have been arranged so that $\circ_i$ is $\geq$ for $i = 1, \ldots, g$; $\circ_i$ is = for $i = g + 1, \ldots, g + e$; $\circ_i$ is $\leq$ for $i = g + e + 1, \ldots, g + e + l = m$.
2. *Adjusted original problem.* Introduce $m + g$ further variables with associated constraints and coefficients for use in f. Thus, replacing j by $j + m$, f becomes $f(x) = \sum_{j=m+1}^{m+n} c_j z_j$ and constraints $z_j \geq 0$ and h_i: $\sum_{j=m+1}^{m+n} a_{ij} z_j \circ_i b_i$, $i = 1, \ldots, m$.
3. *Prepared problem.* For $i = 1, \ldots, g$ replace h_i by H_i: $z_i + \sum_{j=m+1}^{m+n} a_{ij} z_j \; z_{m+n+i} = b_i$, define $c_i = -M (M > 0$ and "large") and $C_{m+n+i} = 0$, and add the constraints $z_i \geq 0$, $z_{m+n+i} \geq 0$. For $i = g + e + 1, \ldots, m$ replace h_i by H_i. $x_i + \sum_{j=m+1}^{m+n} a_{ij} z_j = b_i$, define $c_i = 0$, and adjoin $z_i \geq 0$. Let J run from 1 to $N = n + m + g$; put $Z = [z_j]^T$ and $j = m + 1, \ldots, m + n$. The new function to be maximized is $f(Z) = \sum_{J=1}^{N} c_J z_J$. Actually this is $f(Z) = -M \sum_{i=1}^{g=e} z_i + \sum_{j=m+1}^{vm+n} c_j z_j$, for all other coefficients are zero. Thus for $J = g + e + 1, \ldots, m$ and $m + n + 1, \ldots, N$ the variables z_j make no contribution to f. They are called slack variables, since they take up the slack permitted by the inequalities ($\leq$ and $\geq$) in h_i. Any variable z_j, $i = 1, \ldots, g + e$ whose value is not zero gives rise to a large negative term $-Mz_i$. Such a term will keep $f(Z)$ less than it would be with that $z_i = 0$. The effect of $c_i = -M (i = 1, \ldots, g + e)$ is to make it likely that the optimal solution will have the artificial variables $z_i = 0 (i = 1, \ldots, g + e)$.

The prepared problem now has the form—maximize $f(Z) = \sum_{J=1}^{N} c_J z_J$ subject to $z_J \geq 0$ and H_i: $\sum_{j=1}^{N} a_{ij} z_j = (i = 1, \ldots, m)$, where $b_i \geq 0$,

$$a_{i\beta} = \delta_{i\beta} = \begin{cases} 0 & i \neq \beta \\ 1 & i \neq \beta \end{cases} (\beta = 1, \ldots, m), \; a_{i,m+n+\beta} = -\delta_{i\beta} \; (\beta = 1, \ldots, g)$$

and a_{ij} came from h_i.

The set of feasible points S_P (points satisfying all constraints) for the prepared problem is not empty, and $f(Z)$ is bounded above on S_P.

STATISTICS

General References: 69, 93, 134, 169, 186, 211, 265, 291.

INTRODUCTION

Statistics represents a body of knowledge which enables one to deal with quantitative data reflecting any degree of uncertainty. There are six basic aspects of applied statistics. These are:

1. Type of data
2. Random variables
3. Models
4. Parameters
5. Sample statistics
6. Characterization of chance occurrences

From these can be developed strategies and procedures for dealing with (1) estimation and (2) inferential statistics. The following has been directed more toward inferential statistics because of its broader utility.

Detailed illustrations and examples are used throughout to develop basic statistical methodology for dealing with a broad area of applications. However, in addition to this material, there are many specialized topics as well as some very subtle areas which have not been discussed. The references should be used for more detailed information.

Type of Data In general, statistics deals with two types of data: counts and measurements. Counts represent the number of discrete outcomes, such as the number of defective parts in a shipment, the number of lost-time accidents, and so forth. Measurement data are treated as a continuum. For example, the tensile strength of a synthetic yarn theoretically could be measured to any degree of precision. A subtle aspect associated with count and measurement data is that some types of count data can be dealt with through the application of techniques which have been developed for measurement data alone. This ability is due to the fact that some simplified measurement statistics serve as an excellent approximation for the more tedious count statistics.

Random Variables Applied statistics deals with quantitative data. In tossing a fair coin the successive outcomes would tend to be different, with heads and tails occurring randomly over a period of time. Given a long strand of synthetic fiber, the tensile strength of successive samples would tend to vary significantly from sample to sample.

Counts and measurements are characterized as random variables, that is, observations which are susceptible to chance. Virtually all quantitative data are susceptible to chance in one way or another.

Models Part of the foundation of statistics consists of the mathematical models which characterize an experiment. The models themselves are mathematical ways of describing the probability, or relative likelihood, of observing specified values of random variables. For example, in tossing a coin once, a random variable x could be defined by assigning to x the value 1 for a head and 0 for a tail. Given a fair coin, the probability of observing a head on a toss would be a .5, and similarly for a tail. Therefore, the mathematical model governing this experiment can be written as

x	$P(x)$
0	.5
1	.5

where $P(x)$ stands for what is called a probability function. This term is reserved for count data, in that probabilities can be defined for particular outcomes.

The probability function that has been displayed is a very special case of the more general case, which is called the binomial probability distribution.

For measurement data which are considered continuous, the term *probability density* is used. For example, consider a spinner wheel which conceptually can be thought of as being marked off on the circumference infinitely precisely from 0 up to, but not including, 1. In spinning the wheel, the probability of the wheel's stopping at a specified marking point at any particular x value, where $0 \leq x < 1$, is zero, for example, stopping at the value $x = \sqrt{.5}$. For the spinning wheel, the probability density function would be defined by $f(x) = 1$ for $0 \leq x < 1$. Graphically, this is shown in Fig. 3-55. The relative-probability

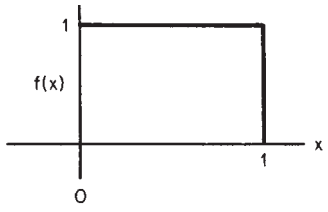

FIG. 3-55 Density function.

concept refers to the fact that density reflects the relative likelihood of occurrence; in this case, each number between 0 and 1 is equally likely.

For measurement data, probability is defined by the area under the curve between specified limits. A density function always must have a total area of 1.

Example For the density of Fig. 3-55 the

$$P[0 \le x \le .4] = .4$$
$$P[.2 \le x \le .9] = .7$$
$$P[.6 \le x < 1] = .4$$

and so forth. Since the probability associated with any particular point value is zero, it makes no difference whether the limit point is defined by a closed interval (≤ or ≥) or an open interval (< or >).

Many different types of models are used as the foundation for statistical analysis. These models are also referred to as **populations.**

Parameters As a way of characterizing probability functions and densities, certain types of quantities called parameters can be defined. For example, the center of gravity of the distribution is defined to be the population mean, which is designated as μ. For the coin toss $\mu = .5$, which corresponds to the average value of x; i.e., for half of the time x will take on a value 0 and for the other half a value 1. The average would be .5. For the spinning wheel, the average value would also be .5.

Another parameter is called the **standard deviation,** which is designated as σ. The square of the standard deviation is used frequently and is called the popular **variance,** σ^2. Basically, the standard deviation is a quantity which measures the spread or dispersion of the distribution from its mean μ. If the spread is broad, then the standard deviation will be larger than if it were more constrained.

For specified probability and density functions, the respective means and variances are defined by the following:

Probability functions	Probability density functions
$E(x) = \mu = \sum_x x\,P(x)$	$E(x) = \mu = \int_x x\,f(x)\,dx$
$\mathrm{Var}(x) = \sigma^2 = \sum_x (x-\mu)^2\,P(x)$	$\mathrm{Var}(x) = \sigma^2 = \int_x (x-\mu)^2\,f(x)\,dx$

where $E(x)$ is defined to be the expected or average value of x.

Sample Statistics Many types of sample statistics will be defined. Two very special types are the **sample mean,** designated as $\bar{x}$, and the sample standard deviation, designated as s. These are, by definition, random variables. Parameters like μ and σ are not random variables; they are fixed constants.

Example In an experiment, six random numbers (rounded to four decimal places) were observed from the uniform distribution $f(x) = 1$ for $0 \le x < 1$:

.1009
.3754
.0842
.9901
.1280
.6606

The sample mean corresponds to the arithmetic average of the observations, which will be designated as x_1 through x_6, where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ with } n = 6$$
$$= (1/6)(.1009 + .3754 + \cdots + .6606)$$
$$= .3899$$

The sample standard deviation s is defined by the computation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$
$$= \sqrt{\frac{n\sum x_i^2 - (\sum x_i)^2}{n(n-1)}}$$

In effect, this represents the root of a statistical average of the squares. The divisor quantity $(n-1)$ will be referred to as the degrees of freedom.

The value of $n-1$ is used in the denominator because the deviations from the sample average must total zero, or

$$\sum (x_i - \bar{x}) = 0$$

Thus knowing $n-1$ values of $x_i - \bar{x}$ permits calculation of the nth value of $x_i - \bar{x}$.

The sample value of the standard deviation for the data given is .3686. The following is a tabulation of the deviations $(x_i - \bar{x}_j)$ for the data:

x	$x - \bar{x}$
.1009	−.2890
.3754	−.0145
.0842	−.3057
.9901	.6002
.1280	−.2619
.6606	.2707
$\bar{x} = .3899$	$s = .3686$

In effect, the standard deviation quantifies the relative magnitude of the deviation numbers, i.e., a special type of "average" of the distance of points from their center. In statistical theory, it turns out that the corresponding variance quantities s^2 have remarkable properties which make possible broad generalities for sample statistics and therefore also their counterparts, the standard deviations.

For the corresponding population, the parameter values are $\mu = .50$ and $\sigma = .2887$. If, instead of using individual observations only, averages of six were reported, then the corresponding population parameter values would be $\mu = .50$ and $\sigma_{\bar{x}} = \sigma/\sqrt{6} = .1179$. The corresponding variance for an average will be written occasionally as $\mathrm{Var}(\bar{x}) = \mathrm{var}(x)/n$. In effect, the variance of an average is inversely proportional to the sample size n, which reflects the fact that sample averages will tend to cluster about μ much more closely than individual observations. This is illustrated in greater detail under "Measurement Data and Sampling Densities."

Characterization of Chance Occurrences To deal with a broad area of statistical applications, it is necessary to characterize the way in which random variables will vary by chance alone. The basic foundation for this characteristic is laid through a density called the gaussian, or normal, distribution.

Determining the area under the normal curve is a very tedious procedure. However, by standardizing a random variable that is normally distributed, it is possible to relate all normally distributed random variables to one table. The standardization is defined by the identity $z = (x - \mu)/\sigma$, where z is called the unit normal. Further, it is possible to standardize the sampling distribution of averages $\bar{x}$ by the identity $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$.

A remarkable property of the normal distribution is that, almost regardless of the distribution of x, sample averages $\bar{x}$ will approach the gaussian distribution as n gets large. Even for relatively small values of n, of about 10, the approximation in most cases is quite close. For example, sample averages of size 10 from the uniform distribution will have essentially a gaussian distribution.

Also, in many applications involving count data, the normal distribution can be used as a close approximation. In particular, the approximation is quite close for the binomial distribution within certain guidelines.

ENUMERATION DATA AND PROBABILITY DISTRIBUTIONS

Introduction Many types of statistical applications are characterized by enumeration data in the form of counts. Examples are the number of lost-time accidents in a plant, the number of defective items in a sample, and the number of items in a sample that fall within several specified categories.

The sampling distribution of count data can be characterized through probability distributions. In many cases, count data are appropriately interpreted through their corresponding distributions. However, in other situations analysis is greatly facilitated through distributions which have been developed for measurement data. Examples of each will be illustrated in the following subsections.

Binomial Probability Distribution

Nature Consider an experiment in which each outcome is classified into one of two categories, one of which will be defined as a success and the other as a failure. Given that the probability of success p is constant from trial to trial, then the probability of observing a specified number of successes x in n trials is defined by the binomial distribution. The sequence of outcomes is called a **Bernoulli process,**

Nomenclature

n = total number of trials
x = number of successes in n trials
p = probability of observing a success on any one trial
$\hat{p} = x/n$, the proportion of successes in n trials

Probability Law

$$P(x) = P\left(\frac{x}{n}\right) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

where $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$

Properties $E(x) = np$ $\quad \text{Var}(x) = np(1-p)$

$E(\hat{p}) = p$ $\quad \text{Var}(\hat{p}) = p(1-p)/n$

Geometric Probability Distribution

Nature Consider an experiment in which each outcome is classified into one of two categories, one of which will be defined as a success and the other as a failure. Given that the probability of success p is constant from trial to trial, then the probability of observing the first success on the xth trial is defined by the geometric distribution.

Nomenclature

p = probability of observing a success on any one trial
x = the number of trials to obtain the first success

Probability Law

$$P(x) = p(1-p)^{x-1} \qquad x = 1, 2, 3, \ldots$$

Properties

$$E(x) = 1/p \qquad \text{Var}(x) = (1-p)/p^2$$

Poisson Probability Distribution

Nature In monitoring a moving threadline, one criterion of quality would be the frequency of broken filaments. These can be identified as they occur through the threadline by a broken-filament detector mounted adjacent to the threadline. In this context, the random occurrences of broken filaments can be modeled by the Poisson distribution. This is called a Poisson process and corresponds to a probabilistic description of the frequency of defects or, in general, what are called arrivals at points on a continuous line or in time. Other examples include:

1. The number of cars (arrivals) that pass a point on a high-speed highway between 10:00 and 11:00 A.M. on Wednesdays
2. The number of customers arriving at a bank between 10:00 and 10:10 A.M.
3. The number of telephone calls received through a switchboard between 9:00 and 10:00 A.M.
4. The number of insurance claims that are filed each week
5. The number of spinning machines that break down during 1 day at a large plant.

Nomenclature

x = total number of arrivals in a total length L or total period T
a = average rate of arrivals for a unit length or unit time
$\lambda = aL$ = expected or average number of arrivals for the total length L
$\lambda = aT$ = expected or average number of arrivals for the total time T

Probability Law Given that a is constant for the total length L or period T, the probability of observing x arrivals in some period L or T is given by

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0, 1, 2, \ldots$$

Properties $E(x) = \lambda$ $\quad \text{Var}(x) = \lambda$

Example The number of broken filaments in a threadline has been averaging .015 per yard. What is the probability of observing exactly two broken filaments in the next 100 yd? In this example, $a = .015$/yd and $L = 100$ yd; therefore $\lambda = (.015)(100) = 1.5$:

$$P(x = 2) = \frac{(1.5)^2}{2!} e^{-1.5} = .2510$$

Example A commercial item is sold in a retail outlet as a unit product. In the past, sales have averaged 10 units per month with no seasonal variation. The retail outlet must order replacement items 2 months in advance. If the outlet starts the next 2-month period with 25 items on hand, what is the probability that it will stock out before the end of the second month?

Given $a = 10$/month, then $\lambda = 10 \times 2 = 20$ for the total period of 2 months:

$$P(x \geq 26) = \sum_{26}^{\infty} P(x) = 1 - \sum_{0}^{25} P(x)$$

$$\sum_{0}^{25} \frac{20^x}{x!} e^{-20} = e^{-20}\left[1 + \frac{20}{1} + \frac{20^2}{2!} + \cdots + \frac{20^{25}}{25!}\right]$$

$$= .887815$$

Therefore $P(x \geq 26) = .112185$ or roughly an 11 percent chance of a stockout.

Hypergeometric Probability Distribution

Nature In an experiment in which one samples from a relatively small group of items, each of which is classified in one of two categories, A or B, the hypergeometric distribution can be defined. One example is the probability of drawing two red and two black cards from a deck of cards. The hypergeometric distribution is the analog of the binomial distribution when successive trials are not independent, i.e., when the total group of items is not infinite. This happens when the drawn items are not replaced.

Nomenclature

N = total group size
n = sample group size
X = number of items in the total group with a specified attribute A
$N - X$ = number of items in the total group with the other attribute B
x = number of items in the sample with a specified attribute A
$n - x$ = number of items in the sample with the other attribute B

	Population	Sample
Category A	X	x
Category B	$N - X$	$n - x$
Total	N	n

Probability Law

$$P(x) = \binom{N-X}{n-x}\binom{X}{x} \Big/ \binom{N}{n}$$

$$E(x) = \frac{nX}{N}$$

$$\text{var}(x) = nP(1-P)\frac{N-n}{N-1}$$

Example What is the probability that an appointed special committee of 4 has no female members when the members are randomly selected from a candidate group of 10 males and 7 females?

$$P(x=0)=\frac{\binom{10}{4}\binom{7}{0}}{\binom{17}{4}}=.0882$$

Example A bin contains 300 items, of which 240 are good and 60 are defective. In a sample of 6 what is the probability of selecting 4 good and 2 defective items by chance?

$$P(x)=\frac{\binom{240}{4}\binom{60}{2}}{\binom{300}{6}}=.2478$$

Multinomial Distribution

Nature For an experiment in which successive outcomes can be classified into two or more categories and the probabilities associated with the respective outcomes remain constant, then the experiment can be characterized through the multinomial distribution.

Nomenclature

n = total number of trials
k = total number of distinct categories
p_j = probability of observing category j on any one trial, $j = 1, 2, \ldots, k$
x_j = total number of occurrences in category j in n trials

Probability Law

$$P(x_1, x_2, \ldots, x_k)=\frac{n!}{x_1!x_2!\ldots x_k!}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}$$

Example In tossing a die 12 times, what is the probability that each face value will occur exactly twice?

$$p(2,2,2,2,2,2)=\frac{12!}{2!2!2!2!2!2!}\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2=.003438$$

MEASUREMENT DATA AND SAMPLING DENSITIES

Introduction The following example data are used throughout this subsection to illustrate concepts. Consider, for the purpose of illustration, that five synthetic-yarn samples have been selected randomly from a production line and tested for tensile strength on each of 20 production days. For this, assume that each group of five corresponds to a day, Monday through Friday, for a period of 4 weeks:

Monday 1	Tuesday 2	Wednesday 3	Thursday 4	Friday 5	Groups of 25 pooled
36.48	38.06	35.28	36.34	36.73	
35.33	31.86	36.58	36.25	37.17	
35.92	33.81	38.81	30.46	33.07	
32.28	30.30	33.31	37.37	34.27	
31.61	35.27	33.88	37.52	36.94	
$\bar{x}$ = 34.32	33.86	35.57	35.59	35.64	35.00
s = 2.22	3.01	2.22	2.92	1.85	2.40
6	7	8	9	10	
38.67	36.62	35.03	35.80	36.82	
32.08	33.05	36.22	33.16	36.49	
33.79	35.43	32.71	35.19	32.83	
32.85	36.63	32.52	32.91	32.43	
35.22	31.46	27.23	35.44	34.16	
$\bar{x}$ = 34.52	34.64	32.74	34.50	34.54	34.19
s = 2.60	2.30	3.46	1.36	2.03	2.35
11	12	13	14	15	
39.63	34.52	36.05	36.64	31.57	
34.38	37.39	35.36	31.18	36.21	
36.51	34.16	35.00	36.13	33.84	
30.00	35.76	33.61	37.51	35.01	
39.64	37.63	36.98	39.05	34.95	
$\bar{x}$ = 36.03	35.89	35.40	36.10	34.32	35.55
s = 4.04	1.59	1.25	2.96	1.75	2.42

Monday 16	Tuesday 17	Wednesday 18	Thursday 19	Friday 20	Groups of 25 pooled
37.68	35.97	33.71	35.61	36.65	
36.38	35.92	32.34	37.13	37.91	
38.43	36.51	33.29	31.37	42.18	
39.07	33.89	32.81	35.89	39.25	
33.06	36.01	37.13	36.33	33.32	
$\bar{x}$ = 36.92	35.66	33.86	35.27	37.86	35.91
s = 2.38	1.02	1.90	2.25	3.27	2.52

Pooled sample of 100: $\bar{x} = 35.16$ $s = 2.47$

Even if the process were at steady state, tensile strength, a key property would still reflect some variation. Steady state, or stable operation of any process, has associated with it a characteristic variation. Superimposed on this is the testing method, which is itself a process with its own characteristic variation. The observed variation is a composite of these two variations.

Assume that the table represents "typical" production-line performance. The numbers themselves have been generated on a computer and represent random observations from a population with $\mu = 35$ and a population standard deviation $\sigma = 2.45$. The sample values reflect the way in which tensile strength can vary by chance alone. In practice, a production supervisor unschooled in statistics but interested in high tensile performance would be despondent on the eighth day and exuberant on the twentieth day. If the supervisor were more concerned with uniformity, the lowest and highest points would have been on the eleventh and seventeenth days.

An objective of statistical analysis is to serve as a guide in decision making in the context of normal variation. In the case of the production supervisor, it is to make a decision, with a high probability of being correct, that something has in fact changed the operation.

Suppose that an engineering change has been made in the process and five new tensile samples have been tested with the results:

36.81
38.34 $\bar{x} = 37.14$
34.87 $s = 1.85$
39.58
36.12

In this situation, management would inquire whether the product has been improved by increased tensile strength. To answer this question, in addition to a variety of analogous questions, it is necessary to have some type of scientific basis upon which to draw a conclusion.

A scientific basis for the evaluation and interpretation of data is contained in the accompanying table descriptions. These tables characterize the way in which sample values will vary by chance alone in the context of individual observations, averages, and variances.

Table number	Designated symbol	Variable	Sampling distribution of
3-4	z	$\frac{x-\mu}{\sigma}$	Observations*
3-4	z	$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$	Averages
3-5	t	$\frac{\bar{x}-\mu}{s/\sqrt{n}}$	Averages when σ is unknown*
3-6	χ^2	$(s^2/\sigma^2)(\text{df})$	Variances*
3-7	F	s_1^2/s_2^2	Ratio of two independent sample variances*

*When sampling from a gaussian distribution.

Normal Distribution of Observations Many types of data follow what is called the gaussian, or bell-shaped, curve; this is especially true of averages. Basically, the gaussian curve is a purely mathematical function which has very special properties. However, owing to some mathematically intractable aspects primary use of the function is restricted to tabulated values.

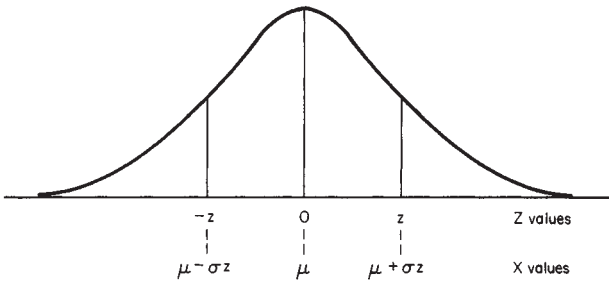

FIG. 3-56 Transformation of z values.

Basically, the tabled values represent area (proportions or probability) associated with a scaling variable designated by Z in Fig. 3-56. The normal curve is centered at 0, and for particular values of Z, designated as z, the tabulated numbers represent the corresponding area under the curve between 0 and z. For example, between 0 and 1 the area is .3413. (Get this number from Table 3-4. The value of A includes the area on both sides of zero. Thus we want $A/2$. For $z = 1$, $A = 0.6827$, $A/2 = 0.3413$. For $z = 2$, $A/2 = 0.4772$.) The area between 0 and 2 is .4772; therefore, the area between 1 and 2 is .4772 − .3413 = .1359.

Also, since the normal curve is symmetric, areas to the left can be determined in exactly the same way. For example, the area between −2 and +1 would include the area between −2 and 0, .4772 (the same as 0 to 2), plus the area between 0 and 1, .3413, or a total area of .8185.

Any types of observation which are applicable to the normal curve can be transformed to Z values by the relationship $z = (x - \mu)/\sigma$ and, conversely, Z values to x values by $x = \mu + \sigma z$, as shown in Fig. 3-56. For example, for tensile strength, with $\mu = 35$ and $\sigma = 2.45$, this would dictate $z = (x - 35)/2.45$ and $x = 35 + 2.45z$.

Example What proportion of tensile values will fall between 34 and 36?

$$z_1 = (34 - 35)/2.45 = -.41 \qquad z_2 = (36 - 35)/2.45 = .41$$

$$P[-.41 \le z \le .41] = .3182, \text{ or roughly 32 percent}$$

The value 0.3182 is interpolated from Table 3-4 using $z = \pm 0.40$, $A = 0.3108$, and $z = \pm 0.45$, $A = 0.3473$.

Example What midrange of tensile values will include 95 percent of the sample values? Since $P[-1.96 \le z \le 1.96] = .95$, the corresponding values of x are

$$x_1 = 35 - 1.96(2.45) = 30.2$$

$$x_2 = 35 + 1.96(2.45) = 39.8$$

or

$$P[30.2 \le x \le 39.8] = .95$$

TABLE 3-4 Ordinates and Areas between Abscissa Values −z and +z of the Normal Distribution Curve

z	X	Y	A	$1 - A$	z	X	Y	A	$1 - A$
0	μ	0.399	0.0000	1.0000	±1.50	μ ± 1.50σ	0.1295	0.8664	0.1336
±0.05	μ ± 0.05σ	.398	.0399	0.9601	±1.55	μ ± 1.55σ	.1200	.8789	.1211
± .10	μ ± .10σ	.397	.0797	.9203	±1.60	μ ± 1.60σ	.1109	.8904	.1096
± .15	μ ± .15σ	.394	.1192	.8808	±1.65	μ ± 1.65σ	.1023	.9011	.0989
± .20	μ ± .20σ	.391	.1585	.8415	±1.70	μ ± 1.70σ	.0940	.9109	.0891
± .25	μ ± .25σ	.387	.1974	.8026	±1.75	μ ± 1.75σ	.0863	.9199	.0801
± .30	μ ± .30σ	.381	.2358	.7642	±1.80	μ ± 1.80σ	.0790	.9281	.0719
± .35	μ ± .35σ	.375	.2737	.7263	±1.85	μ ± 1.85σ	.0721	.9357	.0643
± .40	μ ± .40σ	.368	.3108	.6892	±1.90	μ ± 1.90σ	.0656	.9446	.0574
± .45	μ ± .45σ	.361	.3473	.6527	±1.95	μ ± 1.95σ	.0596	.9488	.0512
± .50	μ ± .50σ	.352	.3829	.6171	±2.00	μ ± 2.00σ	.0540	.9545	.0455
± .55	μ + .55σ	.343	.4177	.5823	±2.05	μ ± 2.05σ	.0488	.9596	.0404
± .60	μ ± .60σ	.333	.4515	.5485	±2.10	μ ± 2.10σ	.0440	.9643	.0357
± .65	μ ± .65σ	.323	.4843	.5157	±2.15	μ ± 2.15σ	.0396	.9684	.0316
± .70	μ ± .70σ	.312	.5161	.4839	±2.20	μ ± 2.20σ	.0335	.9722	.0278
± .75	μ ± .75σ	.301	.5467	.4533	±2.25	μ ± 2.25σ	.0317	.9756	.0244
± .80	μ ± .80σ	.290	.5763	.4237	±2.30	μ ± 2.30σ	.0283	.9786	.0214
± .85	μ ± .85σ	.278	.6047	.3953	±2.35	μ ± 2.35σ	.0252	.9812	.0188
± .90	μ ± .90σ	.266	.6319	.3681	±2.40	μ ± 2.40σ	.0224	.9836	.0164
± .95	μ ± .95σ	.254	.6579	.3421	±2.45	μ ± 2.45σ	.0198	.9857	.0143
±1.00	μ ± 1.00σ	.242	.6827	.3173	±2.50	μ ± 2.50σ	.0175	.9876	.0124
±1.05	μ ± 1.05σ	.230	.7063	.2937	±2.55	μ ± 2.55σ	.0154	.9892	.0108
±1.10	μ ± 1.10σ	.218	.7287	.2713	±2.60	μ ± 2.60σ	.0136	.9907	.0093
±1.15	μ ± 1.15σ	.206	.7499	.2501	±2.65	μ ± 2.65σ	.0119	.9920	.0080
±1.20	μ ± 1.20σ	.194	.7699	.2301	±2.70	μ ± 2.70σ	.0104	.9931	.0069
±1.25	μ ± 1.25σ	.183	.7887	.2113	±2.75	μ ± 2.75σ	.0091	.9940	.0060
±1.30	μ ± 1.30σ	.171	.8064	.1936	±2.80	μ ± 2.80σ	.0079	.9949	.0051
±1.35	μ ± 1.35σ	.160	.8230	.1770	±2.85	μ ± 2.85σ	.0069	.9956	.0044
±1.40	μ ± 1.40σ	.150	.8385	.1615	±2.90	μ ± 2.90σ	.0060	.9963	.0037
±1.45	μ ± 1.45σ	.139	.8529	.1471	±2.95	μ ± 2.95σ	.0051	.9968	.0032
±1.50	μ ± 1.50σ	.130	.8664	.1336	±3.00	μ ± 3.00σ	.0044	.9973	.0027
					±4.00	μ ± 4.00σ	.0001	.99994	.00006
					±5.00	μ ± 5.00s	.000001	.9999994	.0000006
±0.000	μ	0.3989	.0000	1.0000	±1.036	μ ± 1.036σ	0.2331	0.7000	0.3000
± .126	μ ± 0.126σ	.3958	.1000	0.9000	±1.282	μ ± 1.282σ	.1755	.8000	.2000
± .253	μ ± .253σ	.3863	.2000	.8000	±1.645	μ ± 1.645σ	.1031	.9000	.1000
± .385	μ ± .385σ	.3704	.3000	.7000	±1.960	μ + 1.960σ	.0584	.9500	.0500
± .524	μ ± .524σ	.3477	.4000	.6000	±2.576	μ ± 2.576σ	.0145	.9900	.0100
± .674	μ ± .674σ	.3178	.5000	.5000	±3.291	μ ± 3.291σ	.0018	.9990	.0010
± .842	μ ± .842σ	.2800	.6000	.4000	±3.891	μ ± 3.891σ	.0002	.9999	.0001

Normal Distribution of Averages An examination of the tensile-strength data previously tabulated would show that the range (largest minus the smallest) of tensile strength within days averages 5.72. The average range in $\bar{x}$ values within each week is 2.37, while the range in the four weekly averages is 1.72. This reflects the fact that averages tend to be less variable in a predictable way. Given that the variance of x is var $(x) = \sigma^2$, then the variance of $\bar{x}$ based on n observations is var $(\bar{x}) = \sigma^2/n$.

For averages of n observations, the corresponding relationship for the Z-scale relationship is

$$z = (\bar{x} - \mu/\sigma/\sqrt{n}) \qquad \text{or} \qquad \bar{x} = \mu + \frac{\sigma}{\sqrt{n}} z$$

Example What proportion of daily tensile averages will fall between 34 and 36?

$$z_1 = (34 - 35)/(2.45/\sqrt{5}) = -.91 \qquad z_2 = (36 - 35)/(2.45/\sqrt{5}) = .91$$

$$P[-.91 \le z \le .91] = .6372, \text{ or roughly 64 percent}$$

Example What midrange of daily tensile averages will include 95 percent of the sample values?

$$x_1 = 35 - 1.96(2.45/\sqrt{5}) = 32.85$$

$$x_2 = 35 + 1.96(2.45/\sqrt{5}) = 37.15$$

or

$$P[32.85 \le \bar{x} \le 37.15] = .95$$

Example What proportion of weekly tensile averages will fall between 34 and 36?

$$z_1 = (34 - 35)/(2.45/\sqrt{25}) = -2.04$$

$$z_2 = (36 - 35)/(2.45/\sqrt{25}) = 2.04$$

$$P[-2.04 \le z \le 2.04] = .9586, \text{ or roughly 96 percent}$$

Distribution of Averages The normal curve relies on a knowledge of σ, or in special cases, when it is unknown, s can be used with the normal curve as an approximation when $n > 30$. For example, with $n > 30$ the intervals $\bar{x} \pm s$ and $\bar{x} \pm 2s$ will include roughly 68 and 95 percent of the sample values respectively when the distribution is normal.

In applications sample sizes are usually small and σ unknown. In these cases, the t distribution can be used where

$$t = (\bar{x} - \mu)/(s/\sqrt{n}) \qquad \text{or} \qquad \bar{x} = \mu + ts/\sqrt{n}$$

The t distribution is also symmetric and centered at zero. It is said to be robust in the sense that even when the individual observations x are not normally distributed, sample averages of x have distributions which tend toward normality as n gets large. Even for small n of 5 through 10, the approximation is usually relatively accurate.

In reference to the tensile-strength table, consider the summary statistics $\bar{x}$ and s by days. For each day, the t statistic could be computed. If this were repeated over an extensive simulation and the resultant t quantities plotted in a frequency distribution, they would match the corresponding distribution of t values summarized in Table 3-5.

Since the t distribution relies on the sample standard deviation s, the resultant distribution will differ according to the sample size n. To designate this difference, the respective distributions are classified according to what are called the degrees of freedom and abbreviated as df. In simple problems, the df are just the sample size minus 1. In more complicated applications the df can be different. In general, degrees of freedom are the number of quantities minus the number of constraints. For example, four numbers in a square which must have row and column sums equal to zero have only one df, i.e., four numbers minus three constraints (the fourth constraint is redundant).

Example For a sample size $n = 5$, what values of t define a midarea of 90 percent. For 4 df the tabled value of t corresponding to a midarea of 90 percent is 2.132; i.e., $P[-2.132 \le t \le 2.132] = .90$.

Example For a sample size $n = 25$, what values of t define a midarea of 95 percent? For 24 df the tabled value of t corresponding to a midarea of 95 percent is 2.064; i.e., $P[-2.064 \le t \le 2.064] = .95$.

TABLE 3-5 Values of *t*

df	$t_{.40}$	$t_{.30}$	$t_{.20}$	$t_{.10}$	$t_{.05}$	$t_{.025}$	$t_{.01}$	$t_{.005}$
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	0.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921
17	.257	.534	.863	1.333	1.740	2.110	2.567	2.898
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807
24	.256	.531	.857	1.318	1.711	2.064	2.492	2.797
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763
29	.256	.530	.854	1.311	1.699	2.045	2.462	2.756
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576

Above values refer to a single tail outside the indicated limit of t. For example, for 95 percent of the area to be between $-t$ and $+t$ in a two-tailed t distribution, use the values for $t_{0.025}$ or 2.5 percent for each tail.

Example What is the sample value of t for the first day of tensile data?

$$\text{Sample } t = (34.32 - 35)/(2.22/\sqrt{5}) = -.68$$

Note that on the average 90 percent of all such sample values would be expected to fall within the interval ± 2.132.

t Distribution for the Difference in Two Sample Means

Population Variances Are Equal The t distribution can be readily extended to the difference in two sample means when the respective populations have the same variance σ^2:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}}$$

where s_p^2 is a pooled variance defined by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

In this application, the t distribution has $(n_1 + n_2 - 2)$ df.

Population Variances Are Unequal When population variances are unequal, an approximate t quantity can be used:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{a + b}}$$

with

$$a = s_1^2/n_1 \qquad b = s_2^2/n_2$$

and

$$\text{df} = \frac{(a + b)^2}{a^2/(n_1 - 1) + b^2/(n_2 - 1)}$$

Chi-Square Distribution For some industrial applications, product uniformity is of primary importance. The sample standard deviation s is most often used to characterize uniformity. In dealing with this problem, the chi-square distribution can be used where $\chi^2 = (s^2/\sigma^2)$ (df). The chi-square distribution is a family of distributions which are defined by the degrees of freedom associated with the sample variance. For most applications, df is equal to the sample size minus 1.

The probability distribution function is

$$p(y) = y_0 y^{df-2} \exp\left[\frac{-(df)^2}{2}\right]$$

where y_0 is chosen such that the integral of $p(y)$ over all y is one.

In terms of the tensile-strength table previously given, the respective chi-square sample values for the daily, weekly, and monthly figures could be computed. The corresponding df would be 4, 24, and 99 respectively. These numbers would represent sample values from the respective distributions which are summarized in Table 3-6.

In a manner similar to the use of the t distribution, chi square can be interpreted in a direct probabilistic sense corresponding to a midarea of $(1 - \alpha)$:

$$P[\chi_1^2 \le (s^2/\sigma^2)(\text{df}) \le \chi_2^2] = 1 - \alpha$$

where χ_1^2 corresponds to a lower-tail area of $\alpha/2$ and χ_2^2 an upper-tail area of $\alpha/2$.

The basic underlying assumption for the mathematical derivation of chi square is that a random sample was selected from a normal distribution with variance σ^2. When the population is not normal but skewed, square probabilities could be substantially in error.

Example On the basis of a sample size $n = 5$, what midrange of values will include the sample ratio s/σ with a probability of 90 percent?

Use Table 3-6 for 4 df and read $\chi_1^2 = 0.484$ for a lower tail area of 0.05/2, 2.5 percent, and read $\chi_2^2 = 11.1$ for an upper tail area of 97.5 percent.

$$P[.484 \le (s^2/\sigma^2)(4) \le 11.1] = .90$$

or

$$P[.35 \le s/\sigma \le 1.66] = .90$$

Example On the basis of a sample size $n = 25$, what midrange of values will include the sample ratio s/σ with a probability of 90 percent?

$$P[12.4 \le (s^2/\sigma^2)(24) \le 39.4] = .90$$

or

$$P[.72 \le s/\sigma \le 1.28] = .90$$

This states that the sample standard deviation will be at least 72 percent and not more than 128 percent of the population variance 90 percent of the time. Conversely, 10 percent of the time the standard deviation will underestimate or overestimate the population standard deviation by the corresponding amount. Even for samples as large as 25, the relative reliability of a sample standard deviation is poor.

The chi-square distribution can be applied to other types of application which are of an entirely different nature. These include applications which are discussed under "Goodness-of-Fit Test" and "Two-Way Test for Independence of Count Data." In these applications, the mathematical formulation and context are entirely different, but they do result in the same table of values.

***F* Distribution** In reference to the tensile-strength table, the successive pairs of daily standard deviations could be ratioed and squared. These ratios of variance would represent a sample from a distribution called the F distribution or F ratio. In general, the F ratio is defined by the identity

$$F(\gamma_1, \gamma_2) = s_1^2/s_2^2$$

where γ_1 and γ_2 correspond to the respective df's for the sample variances. In statistical applications, it turns out that the primary area of interest is found when the ratios are greater than 1. For this reason, most tabled values are defined for an upper-tail area. However, defining F_2 to be that value corresponding to an upper-tail area of $\alpha/2$, then F_1 for a lower-tail area of $\alpha/2$ can be determined through the identity

TABLE 3-6 Percentiles of the χ^2 Distribution

	Percent									
df	0.5	1	2.5	5	10	90	95	97.5	99	99.5
1	0.000039	0.00016	0.00098	0.0039	0.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

For large values of degrees of freedom the approximate formula

$$\chi_a^2 = n\left(1 - \frac{2}{9n} + z_a\sqrt{\frac{2}{9n}}\right)^3$$

where z_a is the normal deviate and n is the number of degrees of freedom, may be used. For example, $\chi_{.99}^2 = 60[1 - 0.00370 + 2.326(0.06086)]^3 = 60(1.1379)^3 = 88.4$ for the 99th percentile for 60 degrees of freedom.

$$F_1(\gamma_1, \gamma_2) = 1/F_2(\gamma_2, \gamma_1)$$

The F distribution, similar to the chi square, is sensitive to the basic assumption that sample values were selected randomly from a normal distribution.

Example For two sample variances with 4 df each, what limits will bracket their ratio with a midarea probability of 90 percent?

Use Table 3-7 with 4 df in the numerator and denominator and upper 5 percent points (to get both sides totaling 10 percent). The entry is 6.39. Thus:

$$P[1/6.39 \le s_1^2/s_2^2 \le 6.39] = .90$$

or

$$P[.40 \le s_1/s_2 \le 2.53] = .90$$

Confidence Interval for a Mean For the daily sample tensile-strength data with 4 df it is known that $P[-2.132 \le t \le 2.132] = .90$. This states that 90 percent of all samples will have sample t values which fall within the specified limits. In fact, for the 20 daily samples exactly 16 do fall within the specified limits (note that the binomial with $n = 20$ and $p = .90$ would describe the likelihood of exactly none through 20 falling within the prescribed limits—the sample of 20 is only a sample).

Consider the new daily sample (with $n = 5$, $\bar{x} = 37.14$, and $s = 1.85$) which was observed after a process change. In this case, the same probability holds. However, in this instance the sample value of t cannot be computed, since the new μ, under the process change, is not known. Therefore $P[-2.132 \le (37.14 - \mu)/(1.85/\sqrt{5}) \le 2.132] = .90$. In effect, this identity limits the magnitude of possible values for μ. The magnitude of μ can be only large enough to retain the t quantity above -2.132 and small enough to retain the t quantity below $+2.132$. This can be found by rearranging the quantities within the bracket; i.e., $P[35.78 \le \mu \le 38.90] = .90$. This states that we are 90 percent sure that the interval from 35.78 to 38.90 includes the unknown parameter μ.

In general,

$$P\left[\bar{x} - t\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t\frac{s}{\sqrt{n}}\right] = 1 - \alpha$$

where t is defined for an upper-tail area of $\alpha/2$ with $(n - 1)$ df. In this application, the interval limits $(\bar{x} + t\,s/\sqrt{n})$ are random variables which will cover the unknown parameter μ with probability $(1 - \alpha)$. The converse, that we are 100 $(1 - \alpha)$ percent sure that the parameter value is within the interval, is not correct. This statement defines a probability for the parameter rather than the probability for the interval.

Example What values of t define the midarea of 95 percent for weekly samples of size 25, and what is the sample value of t for the second week?

$$P[-2.064 \le t \le 2.064] = .95$$

and

$$(34.19 - 35)/(2.35/\sqrt{25}) = 1.72.$$

Example For the composite sample of 100 tensile strengths, what is the 90 percent confidence interval for μ?

Use Table 3-5 for $t_{.05}$ with df $\approx \infty$.

$$P\left[35.16 - 1.645\frac{2.47}{\sqrt{100}} < \mu < 35.16 + 1.645\frac{2.47}{\sqrt{100}}\right] = .90$$

or

$$P[34.75 \le \mu \le 35.57] = .90$$

Confidence Interval for the Difference in Two Population Means The confidence interval for a mean can be extended to include the difference between two population means. This interval is based on the assumption that the respective populations have the same variance σ^2:

$$(\bar{x}_1 - \bar{x}_2) - ts_p\sqrt{1/n_1 + 1/n_2} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + ts_p\sqrt{1/n_1 + 1/n_2}$$

Example Compute the 95 percent confidence interval based on the original 100-point sample and the subsequent 5-point sample:

$$s_p^2 = \frac{99(2.47)^2 + 4(1.85)^2}{103} = 5.997$$

or

$$s_p = 2.45$$

With 103 df and $\alpha = .05$, $t = \pm 1.96$ using $t_{.025}$ in Table 3-5. Therefore

$$(35.16 - 37.14) \pm 1.96(2.45)\sqrt{1/100 + 1/5} = -1.98 \pm 2.20$$

or

$$-4.18 \le (\mu_1 - \mu_2) \le .22$$

Note that if the respective samples had been based on 52 observations each rather than 100 and 5, the uncertainty factor would have been $\pm .94$ rather than the observed ± 2.20. The interval width tends to be minimum when $n_1 = n_2$.

Confidence Interval for a Variance The chi-square distribution can be used to derive a confidence interval for a population variance σ^2 when the parent population is normally distributed. For a $100(1 - \alpha)$ percent confidence interval

$$\frac{(\text{df})s^2}{\chi_2^2} \le \sigma^2 \le \frac{(\text{df})s^2}{\chi_1^2}$$

where χ_1^2 corresponds to a lower-tail area of $\alpha/2$ and χ_2^2 to an upper-tail area of $\alpha/2$.

Example For the first week of tensile-strength samples compute the 90 percent confidence interval for σ^2 (df = 24, corresponding to $n = 25$, using 5 percent and 95 percent in Table 3-6):

$$\frac{24(2.40)^2}{36.4} \le \sigma^2 \le \frac{24(2.40)^2}{13.8}$$

$$3.80 \le \sigma^2 \le 10.02$$

or

$$1.95 \le \sigma \le 3.17$$

TESTS OF HYPOTHESIS

General Nature of Tests The general nature of tests can be illustrated with a simple example. In a court of law, when a defendant is charged with a crime, the judge instructs the jury initially to presume that the defendant is innocent of the crime. The jurors are then presented with evidence and counterargument as to the defendant's guilt or innocence. If the evidence suggests beyond a reasonable doubt that the defendant did, in fact, commit the crime, they have been instructed to find the defendant guilty; otherwise, not guilty. The burden of proof is on the prosecution.

Jury trials represent a form of decision making. In statistics, an analogous procedure for making decisions falls into an area of statistical inference called **hypothesis testing.**

Suppose that a company has been using a certain supplier of raw materials in one of its chemical processes. A new supplier approaches the company and states that its material, at the same cost, will increase the process yield. If the new supplier has a good reputation, the company might be willing to run a limited test. On the basis of the test results it would then make a decision to change suppliers or not. Good management would dictate that an improvement must be demonstrated (beyond a reasonable doubt) for the new material. That is, the burden of proof is tied to the new material. In setting up a test of hypothesis for this application, the initial assumption would be defined as a null hypothesis and symbolized as H_0. The null hypothesis would state that yield for the new material is no greater than for the conventional material. The symbol μ_0 would be used to designate the known current level of yield for the standard material and μ for the unknown population yield for the new material. Thus, the null hypothesis can be symbolized as H_0: $\mu \le \mu_0$.

The alternative to H_0 is called the alternative hypothesis and is symbolized as H_1: $\mu > \mu_0$.

Given a series of tests with the new material, the average yield $\bar{x}$ would be compared with μ_0. If $\bar{x} < \mu_0$, the new supplier would be dismissed. If $\bar{x} > \mu_0$, the question would be: Is it sufficiently greater in the light of its corresponding reliability, i.e., beyond a reasonable doubt? If the confidence interval for μ included μ_0, the answer would be no, but if it did not include μ_0, the answer would be yes. In this simple application, the formal test of hypothesis would result in the same conclusion as that derived from the confidence interval. However, the utility of tests of hypothesis lies in their generality, whereas confidence intervals are restricted to a few special cases.

TABLE 3-7 *F* Distribution

Upper 5% Points ($F_{.95}$)

Degrees of freedom for denominator	Degrees of freedom for numerator: 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Upper 1% Points ($F_{.99}$)

Degrees of freedom for denominator	Degrees of freedom for numerator: 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5982	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Interpolation should be performed using reciprocals of the degrees of freedom.

Test of Hypothesis for a Mean Procedure

Nomenclature

μ = mean of the population from which the sample has been drawn
σ = standard deviation of the population from which the sample has been drawn
μ_0 = base or reference level
H_0 = null hypothesis
H_1 = alternative hypothesis
α = significance level, usually set at .10, .05, or .01
t = tabled t value corresponding to the significance level α. For a two-tailed test, each corresponding tail would have an area of $\alpha/2$, and for a one-tailed test, one tail area would be equal to α. If σ^2 is known, then z would be used rather than the t.
$t = (\bar{x} - \mu_0)/(s/\sqrt{n})$ = sample value of the test statistic.

Assumptions

1. The n observations $x_1, x_2, \ldots, x_n$ have been selected randomly.
2. The population from which the observations were obtained is normally distributed with an unknown mean μ and standard deviation σ. In actual practice, this is a robust test, in the sense that in most types of problems it is not sensitive to the normality assumption when the sample size is 10 or greater.

Test of Hypothesis

1. Under the null hypothesis, it is assumed that the sample came from a population whose mean μ is equivalent to some base or reference designated by μ_0. This can take one of three forms:

Form 1	Form 2	Form 3
H_0: $\mu = \mu_0$	H_0: $\mu \le \mu_0$	H_0: $\mu \ge \mu_0$
H_1: $\mu \ne \mu_0$	H_1: $\mu > \mu_0$	H_1: $\mu < \mu_0$
Two-tailed test	Upper-tailed test	Lower-tailed test

2. If the null hypothesis is assumed to be true, say, in the case of a two-sided test, form 1, then the distribution of the test statistic t is known. Given a random sample, one can predict how far its sample value of t might be expected to deviate from zero (the midvalue of t) by chance alone. If the sample value of t does, in fact, deviate too far from zero, then this is defined to be sufficient evidence to refute the assumption of the null hypothesis. It is consequently rejected, and the converse or alternative hypothesis is accepted.
3. The rule for accepting H_0 is specified by selection of the α level as indicated in Fig. 3-57. For forms 2 and 3 the α area is defined to be in the upper or the lower tail respectively.
4. The decision rules for each of the three forms are defined as follows: If the sample t falls within the acceptance region, accept H_0 for lack of contrary evidence. If the sample t falls in the critical region, reject H_0 at a significance level of 100α percent.

Example

Application. In the past, the yield for a chemical process has been established at 89.6 percent with a standard deviation of 3.4 percent. A new supplier of raw materials will be used and tested for 7 days.

Procedure

1. The standard of reference is $\mu_0 = 89.6$ with a known $\sigma = 3.4$.
2. It is of interest to demonstrate whether an increase in yield is achieved with the new material; H_0 says it has not; therefore,

$$H_0: \mu \le 89.6 \qquad H_1: \mu > 89.6$$

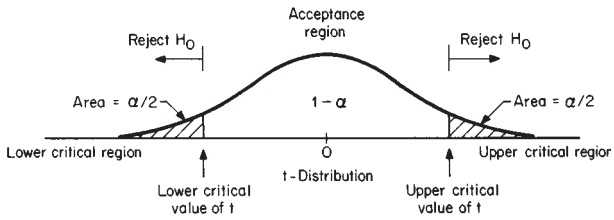

FIG. 3-57 Acceptance region.

3. Select $\alpha = .05$, and since σ is known (the new material would not affect the day-to-day variability in yield), the test statistic would be z with a corresponding critical value $cv(z) = 1.645$ (Table 3-5, df = ∞).
4. The decision rule:

Accept H_0 if sample $z < 1.645$
Reject H_0 if sample $z > 1.645$

5. A 7-day test was carried out, and daily yields averaged 91.6 percent with a sample standard deviation $s = 3.6$ (this is not needed for the test of hypothesis).
6. For the data sample $z = (91.6 - 89.6)/(3.4/\sqrt{7}) = 1.56$.
7. Since the sample $z < cv(z)$, accept the null hypothesis for lack of contrary evidence; i.e., an improvement has not been demonstrated beyond a reasonable doubt.

Example

Application. In the past, the break strength of a synthetic yarn has averaged 34.6 lb. The first-stage draw ratio of the spinning machines has been increased. Production management wants to determine whether the break strength has changed under the new condition.

Procedure

1. The standard of reference is $\mu_0 = 34.6$.
2. It is of interest to demonstrate whether a change has occurred; therefore,

$$H_0: \mu = 34.6 \qquad H_1: \mu \ne 34.6$$

3. Select $\alpha = .05$, and since with the change in draw ratio the uniformity might change, the sample standard deviation would be used, and therefore t would be the appropriate test statistic.
4. A sample of 21 ends was selected randomly and tested on an Instron with the results $\bar{x} = 35.55$ and $s = 2.041$.
5. For 20 df and a two-tailed α level of 5 percent, the critical values of t are given by ± 2.086 with a decision rule (Table 3-5, $t_{.025}$, df = 20):

Accept H_0 if $-2.086 <$ sample $t < 2.086$
Reject H_0 if sample $t < -2.086$ or > 2.086

6. For the data sample $t = (35.55 - 34.6)/(2.041/\sqrt{21}) = 2.133$.
7. Since $2.133 > 2.086$, reject H_0 and accept H_1. It has been demonstrated that an improvement in break strength has been achieved.

Two-Population Test of Hypothesis for Means

Nature Two samples were selected from different locations in a plastic-film sheet and measured for thickness. The thickness of the respective samples was measured at 10 close but equally spaced points in each of the samples. It was of interest to compare the average thickness of the respective samples to detect whether they were significantly different. That is, was there a significant variation in thickness between locations?

From a modeling standpoint statisticians would define this problem as a two-population test of hypothesis. They would define the respective sample sheets as two populations from which 10 sample thickness determinations were measured for each.

In order to compare populations based on their respective samples, it is necessary to have some basis of comparison. This basis is predicated on the distribution of the t statistic. In effect, the t statistic characterizes the way in which two sample means from two separate populations will tend to vary by chance alone when the population means and variances are equal. Consider the following:

Population 1		Population 2	
Normal	Sample 1	Normal	Sample 2
μ_1	n_1	μ_2	n_2
	$\bar{x}_1$		$\bar{x}_2$
σ_1^2	s_1^2	σ_2^2	s_2^2

Consider the hypothesis $\mu_1 = \mu_2$. If, in fact, the hypothesis is correct, i.e., $\mu_1 = \mu_2$ (under the condition $\sigma_1^2 = \sigma_2^2$), then the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is predictable through the t distribution. The observed sample values then can be compared with the corresponding t distribution. If the sample values are reasonably close (as reflected through the α level), that is, $\bar{x}_1$ and $\bar{x}_2$ are not "too different" from each other on the basis of the t distribution, the null hypothesis would be accepted. Conversely, if they deviate from each other "too much" and the deviation is therefore not ascribable to chance, the conjecture would be questioned and the null hypothesis rejected.

Example

Application. Two samples were selected from different locations in a plastic-film sheet. The thickness of the respective samples was measured at 10 close but equally spaced points.

Procedure

1. Demonstrate whether the thicknesses of the respective sample locations are significantly different from each other; therefore,

$$H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2$$

2. Select $\alpha = .05$.
3. Summarize the statistics for the respective samples:

Sample 1		Sample 2	
1.473	1.367	1.474	1.417
1.484	1.276	1.501	1.448
1.484	1.485	1.485	1.469
1.425	1.462	1.435	1.474
1.448	1.439	1.348	1.452
$\bar{x}_1 = 1.434$	$s_1 = .0664$	$\bar{x}_2 = 1.450$	$s_2 = .0435$

4. As a first step, the assumption for the standard t test, that $\sigma_1^2 = \sigma_2^2$, can be tested through the F distribution. For this hypothesis, H_0: $\sigma_1^2 = \sigma_2^2$ would be tested against H_1: $\sigma_1^2 \neq \sigma_2^2$. Since this is a two-tailed test and conventionally only the upper tail for F is published, the procedure is to use the largest ratio and the corresponding ordered degrees of freedom. This achieves the same end result through one table. However, since the largest ratio is arbitrary, it is necessary to define the true α level as twice the value of the tabled value. Therefore, by using Table 3-7 with $\alpha = .05$ the corresponding critical value for $F(9,9) = 3.18$ would be for a true $\alpha = .10$. For the sample,

$$\text{Sample } F = (.0664/.0435)^2 = 2.33$$

Therefore, the ratio of sample variances is no larger than one might expect to observe when in fact $\sigma_1^2 = \sigma_2^2$. There is not sufficient evidence to reject the null hypothesis that $\sigma_1^2 = \sigma_2^2$.

5. For 18 df and a two-tailed α level of 5 percent the critical values of t are given by ± 2.101 (Table 3-5, $t_{0.025}$, $df = 18$).
6. The decision rule:

Accept H_0 if $-2.101 \leq$ sample $t \leq 2.101$
Reject H_0 otherwise

7. For the sample the pooled variance estimate is given by

$$s_p^2 = \frac{9(.0664)^2 + 9(.0435)^2}{9 + 9} = \frac{(.0664)^2 + (.0435)^2}{2} = .00315$$

or

$$s_p = .056$$

8. The sample statistic value of t is

$$\text{Sample } t = \frac{1.434 - 1.450}{.056\sqrt{1/10 + 1/10}} = -.64$$

9. Since the sample value of t falls within the acceptance region, accept H_0 for lack of contrary evidence; i.e., there is insufficient evidence to demonstrate that thickness differs between the two selected locations.

Test of Hypothesis for Paired Observations

Nature In some types of applications, associated pairs of observations are defined. For example, (1) pairs of samples from two populations are treated in the same way, or (2) two types of measurements are made on the same unit. For applications of this type, it is not only more effective but necessary to define the random variable as the difference between the pairs of observations. The difference numbers can then be tested by the standard t distribution.

Examples of the two types of applications are as follows:

1. *Sample treatment*

a. Two types of metal specimens buried in the ground together in a variety of soil types to determine corrosion resistance

b. Wear-rate test with two different types of tractor tires mounted in pairs on n tractors for a defined period of time

2. *Same unit*

a. Blood-pressure measurements made on the same individual before and after the administration of a stimulus

b. Smoothness determinations on the same film samples at two different testing laboratories

Test of Hypothesis for Matched Pairs: Procedure

Nomenclature

d_i = sample difference between the ith pair of observations
s = sample standard deviation of differences
μ = population mean of differences
σ = population standard deviation of differences
μ_0 = base or reference level of comparison
H_0 = null hypothesis
H_1 = alternative hypothesis
α = significance level
t = tabled value with $(n - 1)$ df
$t = (\bar{d} - \mu_0)/(s/\sqrt{n})$, the sample value of t

Assumptions

1. The n pairs of samples have been selected and assigned for testing in a random way.
2. The population of differences is normally distributed with a mean μ and variance σ^2. As in the previous application of the t distribution, this is a robust procedure, i.e., not sensitive to the normality assumption if the sample size is 10 or greater in most situations.

Test of Hypothesis

1. Under the null hypothesis, it is assumed that the sample came from a population whose mean μ is equivalent to some base or reference level designated by μ_0. For most applications of this type, the value of μ_0 is defined to be zero; that is, it is of interest generally to demonstrate a difference not equal to zero. The hypothesis can take one of three forms:

Form 1	Form 2	Form 3
H_0: $\mu = \mu_0$	H_0: $\mu \leq \mu_0$	H_0: $\mu \geq \mu_0$
H_1: $\mu \neq \mu_0$	H_1: $\mu > \mu_0$	H_1: $\mu < \mu_0$
Two-tailed test	Upper-tailed test	Lower-tailed test

2. If the null hypothesis is assumed to be true, say, in the case of a lower-tailed test, form 3, then the distribution of the test statistic t is known under the null hypothesis that limits $\mu = \mu_0$. Given a random sample, one can predict how far its sample value of t might be expected to deviate from zero by chance alone when $\mu = \mu_0$. If the sample value of t is too small, as in the case of a negative value, then this would be defined as sufficient evidence to reject the null hypothesis.
3. Select α.
4. The critical values or value of t would be defined by the tabled value of t with $(n - 1)$ df corresponding to a tail area of α. For a two-tailed test, each tail area would be $\alpha/2$, and for a one-tailed test there would be an upper-tail or a lower-tail area of α corresponding to forms 2 and 3 respectively.
5. The decision rule for each of the three forms would be to reject the null hypothesis if the sample value of t fell in that area of the t distribution defined by α, which is called the critical region. Otherwise, the alternative hypothesis would be accepted for lack of contrary evidence.

Example

Application. Pairs of pipes have been buried in 11 different locations to determine corrosion on nonbituminous pipe coatings for underground use. One type includes a lead-coated steel pipe and the other a bare steel pipe.

Procedure

1. The standard of reference is taken as $\mu_0 = 0$, corresponding to no difference in the two types.
2. It is of interest to demonstrate whether either type of pipe has a greater corrosion resistance than the other. Therefore,

$$H_0: \mu = 0 \qquad H_1: \mu \neq 0$$

3. Select $\alpha = .05$. Therefore, with $n = 11$ the critical values of t with 10 df are defined by $t = \pm 2.228$ (Table 3.5, $t_{.025}$).
4. The decision rule:

Accept H_0 if $-2.228 \leq$ sample $t \leq 2.228$
Reject H_0 otherwise

5. The sample of 11 pairs of corrosion determinations and their differences are as follows:

Soil type	Lead-coated steel pipe	Bare steel pipe	d = difference
A	27.3	41.4	–14.1
B	18.4	18.9	–0.5
C	11.9	21.7	–9.8
D	11.3	16.8	–5.5
E	14.8	9.0	5.8
F	20.8	19.3	1.5
G	17.9	32.1	–14.2
H	7.8	7.4	0.4
I	14.7	20.7	–6.0
J	19.0	34.4	–15.4
K	65.3	76.2	–10.9

6. The sample statistics:

$$\bar{d} = -6.245 \qquad s^2 = \frac{11\sum d^2 - (\sum d)^2}{11 \times 10} = 52.56$$

or

$$s = 7.25$$

$$\text{Sample } t = (-6.245 - 0)/(7.25/\sqrt{11})$$

$$= -2.86$$

7. Since the sample t of –2.86 < tabled t of –2.228, reject H_0 and accept H_1; that is, it has been demonstrated that, on the basis of the evidence, lead-coated steel pipe has a greater corrosion resistance than bare steel pipe.

Example

Application. A stimulus was tested for its effect on blood pressure. Ten men were selected randomly, and their blood pressure was measured before and after the stimulus was administered. It was of interest to determine whether the stimulus had caused a significant increase in the blood pressure.

Procedure

1. The standard of reference was taken as $\mu_0 \leq 0$, corresponding to no increase.
2. It was of interest to demonstrate an increase in blood pressure if in fact an increase did occur. Therefore,

$$H_0: \mu_0 \leq 0 \qquad H_1: \mu_0 > 0$$

3. Select $\alpha = .05$. Therefore, with $n = 10$ the critical value of t with 9 df is defined by $t = 1.833$ (Table 3-5, $t_{.05}$, one-sided).
4. The decision rule:

Accept H_0 if sample $t < 1.833$

Reject H_0 if sample $t > 1.833$

5. The sample of 10 pairs of blood pressure and their differences were as follows:

Individual	Before	After	d = difference
1	138	146	8
2	116	118	2
3	124	120	–4
4	128	136	8
5	155	174	19
6	129	133	4
7	130	129	–1
8	148	155	7
9	143	148	5
10	159	155	–4

6. The sample statistics:

$$\bar{d} = 4.4 \qquad s = 6.85$$

$$\text{Sample } t = (4.4 - 0)/(6.85/\sqrt{10}) = 2.03$$

7. Since the sample $t = 2.03$ > critical $t = 1.833$, reject the null hypothesis. It has been demonstrated that the population of men from which the sample was drawn tend, as a whole, to have an increase in blood pressure after the stimulus has been given. The distribution of differences d seems to indicate that the degree of response varies by individuals.

Test of Hypothesis for a Proportion

Nature Some types of statistical applications deal with counts and proportions rather than measurements. Examples are (1) the proportion of workers in a plant who are out sick, (2) lost-time worker accidents per month, (3) defective items in a shipment lot, and (4) preference in consumer surveys.

The procedure for testing the significance of a sample proportion follows that for a sample mean. In this case, however, owing to the nature of the problem the appropriate test statistic is Z. This follows from the fact that the null hypothesis requires the specification of the goal or reference quantity p_0, and since the distribution is a binomial proportion, the associated variance is $[p_0(1 - p_0)]n$ under the null hypothesis. The primary requirement is that the sample size n satisfy normal approximation criteria for a binomial proportion, roughly $np > 5$ and $n(1 - p) > 5$.

Test of Hypothesis for a Proportion: Procedure

Nomenclature

p = mean proportion of the population from which the sample has been drawn
p_0 = base or reference proportion
$[p_0(1 - p_0)]/n$ = base or reference variance
$\hat{p} = x/n$ = sample proportion, where x refers to the number of observations out of n which have the specified attribute
H_0 = assumption or null hypothesis regarding the population proportion
H_1 = alternative hypothesis
α = significance level, usually set at .10, .05, or .01
z = Tabled Z value corresponding to the significance level α. The sample sizes required for the z approximation according to the magnitude of p_0 are given in Table 3-5.
$z = (\hat{p} - p_0)/\sqrt{p_0(1 - p_0)/n}$, the sample value of the test statistic

Assumptions

1. The n observations have been selected randomly.
2. The sample size n is sufficiently large to meet the requirement for the Z approximation.

Test of Hypothesis

1. Under the null hypothesis, it is assumed that the sample came from a population with a proportion p_0 of items having the specified attribute. For example, in tossing a coin the population could be thought of as having an unbounded number of potential tosses. If it is assumed that the coin is fair, this would dictate $p_0 = 1/2$ for the proportional number of heads in the population. The null hypothesis can take one of three forms:

Form 1	Form 2	Form 3
$H_0: p = p_0$	$H_0: p \leq p_0$	$H_0: p \geq p_0$
$H_1: p \neq p_0$	$H_1: p > p_0$	$H_1: p < p_0$
Two-tailed test	Upper-tailed test	Lower-tailed test

2. If the null hypothesis is assumed to be true, then the sampling distribution of the test statistic Z is known. Given a random sample, it is possible to predict how far the sample proportion x/n might deviate from its assumed population proportion p_0 through the Z distribution. When the sample proportion deviates too far, as defined by the significance level α, this serves as the justification for rejecting the assumption, that is, rejecting the null hypothesis.
3. The decision rule is given by

Form 1: Accept H_0 if lower critical z < sample z < upper critical z
Reject H_0 otherwise

Form 2: Accept H_0 if sample z < upper critical z
Reject H_0 otherwise

Form 3: Accept H_0 if lower critical z < sample z
Reject H_0 otherwise

Example

Application. A company has received a very large shipment of rivets. One product specification required that no more than 2 percent of the rivets have diameters greater than 14.28 mm. Any rivet with a diameter greater than this would be classified as defective. A random sample of 600 was selected and

tested with a go–no-go gauge. Of these, 16 rivets were found to be defective. Is this sufficient evidence to conclude that the shipment contains more than 2 percent defective rivets?

Procedure

1. The quality goal is $p \le .02$. It would be assumed initially that the shipment meets this standard; i.e., H_0: $p \le .02$.
2. The assumption in step 1 would first be tested by obtaining a random sample. Under the assumption that $p \le .02$, the distribution for a sample proportion would be defined by the z distribution. This distribution would define an upper bound corresponding to the upper critical value for the sample proportion. It would be unlikely that the sample proportion would rise above that value if, in fact, $p \le .02$. If the observed sample proportion exceeds that limit, corresponding to what would be a very unlikely chance outcome, this would lead one to question the assumption that $p \le .02$. That is, one would conclude that the null hypothesis is false. To test, set

$$H_0: p \le .02 \qquad H_1: p > .02$$

3. Select $\alpha = .05$.
4. With $\alpha = .05$, the upper critical value of $Z = 1.645$ (Table 3-5, $t_{.05}$, df = ∞, one-sided).
5. The decision rule:

$$\text{Accept } H_0 \text{ if sample } z < 1.645$$

$$\text{Reject } H_0 \text{ if sample } z > 1.645$$

6. The sample z is given by

$$\text{Sample } z = \frac{(16/600) - .02}{\sqrt{(.02)(.98)/600}}$$

$$= 1.17$$

7. Since the sample $z < 1.645$, accept H_0 for lack of contrary evidence; there is not sufficient evidence to demonstrate that the defect proportion in the shipment is greater than 2 percent.

Test of Hypothesis for Two Proportions

Nature In some types of engineering and management-science problems, we may be concerned with a random variable which represents a proportion, for example, the proportional number of defective items per day. The method described previously relates to a single proportion. In this subsection two proportions will be considered.

A certain change in a manufacturing procedure for producing component parts is being considered. Samples are taken by using both the existing and the new procedures in order to determine whether the new procedure results in an improvement. In this application, it is of interest to demonstrate statistically whether the population proportion p_2 for the new procedure is less than the population proportion p_1 for the old procedure on the basis of a sample of data.

Test of Hypothesis for Two Proportions: Procedure

Nomenclature

p_1 = population 1 proportion
p_2 = population 2 proportion
n_1 = sample size from population 1
n_2 = sample size from population 2
x_1 = number of observations out of n_1 that have the designated attribute
x_2 = number of observations out of n_2 that have the designated attribute
$\hat{p}_1 = x_1/n_1$, the sample proportion from population 1
$\hat{p}_2 = x_2/n_2$, the sample proportion from population 2
α = significance level
H_0 = null hypothesis
H_1 = alternative hypothesis
z = tabled Z value corresponding to the stated significance level α

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}}, \text{ the sample value of } Z$$

Assumptions

1. The respective two samples of n_1 and n_2 observations have been selected randomly.
2. The sample sizes n_1 and n_2 are sufficiently large to meet the requirement for the Z approximation; i.e., $x_1 > 5$, $x_2 > 5$.

Test of Hypothesis

1. Under the null hypothesis, it is assumed that the respective two samples have come from populations with equal proportions $p_1 = p_2$. Under this hypothesis, the sampling distribution of the corresponding Z statistic is known. On the basis of the observed data, if the resultant sample value of Z represents an unusual outcome, that is, if it falls within the critical region, this would cast doubt on the assumption of equal proportions. Therefore, it will have been demonstrated statistically that the population proportions are in fact not equal. The various hypotheses can be stated:

Form 1	Form 2	Form 3
H_0: $p_1 = p_2$	H_0: $p_1 \le p_2$	H_0: $p_1 \ge p_2$
H_1: $p_1 \ne p_2$	H_1: $p_1 > p_2$	H_1: $p_1 < p_2$
Two-tailed test	Upper-tailed test	Lower-tailed test

2. The decision rule for form 1 is given by
   Accept H_0 if lower critical z < sample z < upper critical z
   Reject H_0 otherwise

Example

Application. A change was made in a manufacturing procedure for component parts. Samples were taken during the last week of operations with the old procedure and during the first week of operations with the new procedure. Determine whether the proportional numbers of defects for the respective populations differ on the basis of the sample information.

Procedure

1. The hypotheses are

$$H_0: p_1 = p_2 \qquad H_1: p_1 \ne p_2$$

2. Select $\alpha = .05$. Therefore, the critical values of z are ± 1.96 (Table 3-4, $A = 0.9500$).
3. For the samples, 75 out of 1720 parts from the previous procedure and 80 out of 2780 parts under the new procedure were found to be defective; therefore,

$$\hat{p}_1 = 75/1720 = .0436 \qquad \hat{p}_2 = 80/2780 = .0288$$

4. The decision rule:

$$\text{Accept } H_0 \text{ if } -1.96 \le \text{sample } Z \le 1.96$$
$$\text{Reject } H_0 \text{ otherwise}$$

5. The sample statistic:

$$\text{Sample } z = \frac{.0436 - .0288}{\sqrt{(.0436)(.9564)/1720 + (.0288)(.9712)/2780}}$$

$$= 2.53$$

6. Since the sample z of 2.53 > tabled z of 1.96, reject H_0 and conclude that the new procedure has resulted in a reduced defect rate.

Goodness-of-Fit Test

Nature A standard die has six sides numbered from 1 to 6. If one were really interested in determining whether a particular die was well balanced, one would have to carry out an experiment. To do this, it might be decided to count the frequencies of outcomes, 1 through 6, in tossing the die N times. On the assumption that the die is perfectly balanced, one would expect to observe $N/6$ occurrences each for 1, 2, 3, 4, 5, and 6. However, chance dictates that exactly $N/6$ occurrences each will not be observed. For example, given a perfectly balanced die, the probability is only 1 chance in 65 that one will observe 1 outcome each, for 1 through 6, in tossing the die 6 times. Therefore, an outcome different from 1 occurrence each can be expected. Conversely, an outcome of six 3s would seem to be too unusual to have occurred by chance alone.

Some industrial applications involve the concept outlined here. The basic idea is to test whether or not a group of observations follows a preconceived distribution. In the case cited, the distribution is uniform; i.e., each face value should *tend* to occur with the same frequency.

Goodness-of-Fit Test: Procedure

Nomenclature Each experimental observation can be classified into one of r possible categories or cells.

r = total number of cells
O_j = number of observations occurring in cell j
E_j = expected number of observations for cell j based on the preconceived distribution
N = total number of observations
f = degrees of freedom for the test. In general, this will be equal to $(r-1)$ minus the number of statistical quantities on which the E_j's are based (see the examples which follow for details).

Assumptions

1. The observations represent a sample selected randomly from a population which has been specified.
2. The number of expectation counts E_j within each category should be roughly 5 or more. If an E_j count is significantly less than 5, that cell should be pooled with an adjacent cell.

Computation for $\mathbf{E_j}$ On the basis of the specified population, the probability of observing a count in cell j is defined by p_j. For a sample of size N, corresponding to N total counts, the expected frequency is given by $E_j = Np_j$.

Test Statistics: Chi Square

$$\chi^2 = \sum_{j=1}^{r} \frac{(O_j - E_j)^2}{E_j} \qquad \text{with } f \text{ df}$$

Test of Hypothesis

1. H_0: The sample came from the specified theoretical distribution
   H_1: The sample did not come from the specified theoretical distribution
2. For a stated level of α,
   Reject H_0 if sample $\chi^2 >$ tabled χ^2
   Accept H_0 if sample $\chi^2 <$ tabled χ^2

Example

Application A production-line product is rejected if one of its characteristics does not fall within specified limits. The standard goal is that no more than 2 percent of the production should be rejected.

Computation

1. Of 950 units produced during the day, 28 units were rejected.
2. The hypotheses:
   H_0: the process is in control
   H_1: the process is not in control
3. Assume that $\alpha = .05$; therefore, the critical value of $\chi^2(1) = 3.84$ (Table 3-6, 95 percent, df = 1). One degree of freedom is defined since $(r-1) = 1$, and no statistical quantities have been computed for the data.
4. The decision rule:
   Reject H_0 if sample $\chi^2 > 3.84$
   Accept H_0 otherwise
5. Since it is assumed that $p = .02$, this would dictate that in a sample of 950 there would be on the average $(.02)(950) = 19$ defective items and 931 acceptable items:

Category	Observed O_j	Expectation $E_j = 950p_j$
Acceptable	922	931
Not acceptable	28	19
Total	950	950

$$\text{Sample } \chi^2 = \frac{(922-931)^2}{931} + \frac{(28-19)^2}{19}$$
$$= 4.35 \text{ with critical } \chi^2 = 3.84$$

6. Conclusion. Since the sample value exceeds the critical value, it would be concluded that the process is not in control.

Example

Application A frequency count of workers was tabulated according to the number of defective items that they produced. An unresolved question is whether the observed distribution is a Poisson distribution. That is, do observed and expected frequencies agree within chance variation?

Computation

1. The hypotheses:
   H_0: there are no significant differences, in number of defective units, between workers
   H_1: there are significant differences
2. Assume that $\alpha = .05$.
3. Test statistic:

No. of defective units	O_j		E_j	
0	8	10	2.06	8.70 pool
1	7		6.64	
2	9		10.73	
3	12		11.55	
4	9		9.33	
5	6		6.03	
6	3	6	3.24	5.66 pool
7	2		1.50	
8	0		.60	
9	1		.22	
≥10	0		.10	
Sum	52		52	

The expectation numbers E_j were computed as follows: For the Poisson distribution, $\lambda = E(x)$; therefore, an estimate of λ is the average number of defective units per worker, i.e., $\lambda = (1/52)(0\times 3 + 1\times 7 + \cdots + 9\times 1) = 3.23$. Given this approximation, the probability of no defective units for a worker would be $(3.23)^0/0!)e^{-3.23} = .0396$. For the 52 workers, the number of workers producing no defective units would have an expectation $E = 52(0.0396) = 2.06$, and so forth.

The sample chi-square value is computed from

$$\chi^2 = \frac{(10-8.70)^2}{8.70} + \frac{(9-10.73)^2}{10.73} + \cdots + \frac{(6-5.66)^2}{5.66}$$
$$= .53$$

4. The critical value of χ^2 would be based on four degrees of freedom. This corresponds to $(r-1)-1$, since one statistical quantity λ was computed from the sample and used to derive the expectation numbers.
5. The critical value of $\chi^2(4) = 9.49$ (Table 3-6) with $\alpha = .05$; therefore, accept H_0.

Two-Way Test for Independence for Count Data

Nature When individuals or items are observed and classified according to two different criteria, the resultant counts can be statistically analyzed. For example, a market survey may examine whether a new product is preferred and if it is preferred due to a particular characteristic.

Count data, based on a random selection of individuals or items which are classified according to two different criteria, can be statistically analyzed through the χ^2 distribution. The purpose of this analysis is to determine whether the respective criteria are dependent. That is, is the product preferred because of a particular characteristic?

Two-Way Test for Independence for Count Data: Procedure

Nomenclature

1. Each observation is classified into each of two categories:
   a. The first one into 2, 3, . . . , or r categories
   b. The second one into 2, 3, . . . , or c categories
2. O_{ij} = number of observations (observed counts) in cell (i, j) with
   $i = 1, 2, \ldots, r$
   $j = 1, 2, \ldots, c$
3. N = total number of observations
4. E_{ij} = computed number for cell (i,j) which is an expectation based on the assumption that the two characteristics are independent
5. R_i = subtotal of counts in row i
6. C_j = subtotal of counts in column j
7. α = significance level
8. H_0 = null hypothesis
9. H_1 = alternative hypothesis
10. χ^2 = critical value of χ^2 corresponding to the significance level α and $(r-1)(c-1)$ df
11. $$\text{Sample } \chi^2 = \sum_{i,j}^{c,r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Assumptions

1. The observations represent a sample selected randomly from a large total population.

2. The number of expectation counts E_{ij} within each cell should be approximately 2 or more for arrays 3×3 or larger. If any cell contains a number smaller than 2, appropriate rows or columns should be combined to increase the magnitude of the expectation count. For arrays 2×2, approximately 4 or more are required. If the number is less than 4, the exact Fisher test should be used.

Test of Hypothesis Under the null hypothesis, the classification criteria are assumed to be independent, i.e.,

H_0: the criteria are independent

H_1: the criteria are not independent

For the stated level of α,

Reject H_0 if sample $\chi^2 >$ tabled χ^2

Accept H_0 otherwise

Computation for $\mathbf{E_{ij}}$ Compute E_{ij} across rows or down columns by using either of the following identities:

$$E_{ij} = C_j \left(\frac{R_i}{N}\right) \text{ across rows}$$

$$E_{ij} = R_i \left(\frac{C_j}{N}\right) \text{ down columns}$$

Sample χ^2 Value

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

In the special case of $r = 2$ and $c = 2$, a more accurate and simplified formula which does not require the direct computation of E_{ij} can be used:

$$\chi^2 = \frac{[|O_{11}O_{22} - O_{12}O_{21}| - \frac{1}{2}N]^2 N}{R_1 R_2 C_1 C_2}$$

Example

Application A market research study was carried out to relate the subjective "feel" of a consumer product to consumer preference. In other words, is the consumer's preference for the product associated with the feel of the product, or is the preference independent of the product feel?

Procedure

1. It was of interest to demonstrate whether an association exists between feel and preference; therefore, assume

H_0: feel and preference are independent

H_1: they are not independent

2. A sample of 200 people was asked to classify the product according to two criteria:

a. Liking for this product

b. Liking for the feel of the product

		Like feel		
		Yes	No	R_i
Like product	Yes	114	13	= 127
	No	55	18	= 73
	C_j	169	31	200

3. Select $\alpha = .05$; therefore, with $(r - 1)(c - 1) = 1$ df, the critical value of χ^2 is 3.84 (Table 3-6, 95 percent).

4. The decision rule:

Accept H_0 if sample $\chi^2 < 3.84$

Reject H_0 otherwise

5. The sample value of χ^2 by using the special formula is

$$\text{Sample } \chi^2 = \frac{[|114 \times 18 - 13 \times 55| - 100]^2 200}{(169)(31)(127)(73)}$$

$$= 6.30$$

6. Since the sample χ^2 of 6.30 > tabled χ^2 of 3.84, reject H_0 and accept H_1. The relative proportionality of $E_{11} = 169(127/200) = 107.3$ to the observed 114 compared with $E_{22} = 31(73/200) = 11.3$ to the observed 18 suggests that when the consumer likes the feel, the consumer tends to like the product, and conversely for not liking the feel. The proportions 169/200 = 84.5 percent and 127/200 = 63.5 percent suggest further that there are other attributes of the product which tend to nullify the beneficial feel of the product.

LEAST SQUARES

When experimental data is to be fit with a mathematical model, it is necessary to allow for the fact that the data has errors. The engineer is interested in finding the parameters in the model as well as the uncertainty in their determination. In the simplest case, the model is a linear equation with only two parameters, and they are found by a least-squares minimization of the errors in fitting the data. Multiple regression is just linear least squares applied with more terms. Nonlinear regression allows the parameters of the model to enter in a nonlinear fashion. The following description of maximum likelihood applies to both linear and nonlinear least squares (Ref. 231). If each measurement point y_i has a measurement error Δy_i that is independently random and distributed with a normal distribution about the true model $y(x)$ with standard deviation σ_i, then the probability of a data set is

$$P = \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2} \left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2 \right] \Delta y \right\}$$

Here, y_i is the measured value, σ_i is the standard deviation of the ith measurement, and Δy is needed to say a measured value $\pm\Delta y$ has a certain probability. Given a set of parameters (maximizing this function), the probability that this data set plus or minus Δy could have occurred is P. This probability is maximized (giving the maximum likelihood) if the negative of the logarithm is minimized.

$$\sum_{i=1}^{N} \left(\frac{y_i - y(x_i)}{\sqrt{2}\,\sigma_i} \right)^2 - N \log \Delta y$$

Since N, σ_i, and Δy are constants, this is the same as minimizing χ^2.

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - y(x_i; a_1, \ldots, a_M)}{\sigma_i} \right]^2$$

with respect to the parameters $\{a_j\}$. Note that the standard deviations $\{\sigma_i\}$ of the measurements are expected to be known. The goodness of fit is related to the number of degrees of freedom, $\nu = N - M$. The probability that χ^2 would exceed a particular value $(\chi_0)^2$ is

$$P = 1 - Q\left(\frac{\nu}{2}, \frac{1}{2} \chi_0^2 \right)$$

where $Q(a, x)$ is the incomplete gamma function

$$Q(a, x) = \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1}\, dt \qquad (a > 0)$$

and $\Gamma(a)$ is the gamma function

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t}\, dt$$

Both functions are tabulated in mathematical handbooks (Ref. 1). The function P gives the goodness of fit. Call χ_0^2 the value of χ^2 at the minimum. Then $P > 0.1$ represents a believable fit; if $Q > 0.001$, it might be an acceptable fit; smaller values of Q indicate the model may be in error (or the σ_i are really larger.) A "typical" value of χ^2 for a moderately good fit is $\chi^2 \sim \nu$. Asymptotically for large ν, the statistic χ^2 becomes normally distributed with a mean ν and a standard deviation $\sqrt{(2\nu)}$ (Ref. 231).

If values σ_i are not known in advance, assume $\sigma_i = \sigma$ (so that its value does not affect the minimization of χ^2). Find the parameters by minimizing χ^2 and compute:

$$\sigma^2 = \sum_{i=1}^{N} \frac{[y_i - y(x_i)]^2}{N}$$

This gives some information about the errors (i.e., the variance and standard deviation of each data point), although the goodness of fit, P, cannot be calculated.

The minimization of χ^2 requires

$$\sum_{i=1}^{N} \left[\frac{y_i - y(x_i)}{\sigma_i^2} \right] \frac{\partial y(x_i; a_1, \ldots, a_M)}{\partial a_k} = 0, \quad k = 1, \ldots, M$$

Linear Least Squares When the model is a straight line

$$\chi^2(a, b) = \sum_{i=1}^{N} \left[\frac{y_i - a - bx_i}{\sigma_i} \right]^2$$

Define $$S = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}, \quad S_x = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}, \quad S_y = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$$

$$S_{xx} = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}, \quad S_{xy} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}, \quad t_i = \frac{1}{\sigma_i}\left(x_i - \frac{S_x}{S}\right), \quad S_{tt} = \sum_{i=1}^{N} t_i^2$$

Then $$b = \frac{1}{S_{tt}} \sum_{i=1}^{N} \frac{t_i y_i}{\sigma_i}, \quad a = \frac{S_y - S_x b}{S}, \quad \sigma_a^2 = \frac{1}{S}\left(1 + \frac{S_x^2}{S S_{tt}}\right), \quad \sigma_b^2 = \frac{1}{S_{tt}}$$

$$\text{Cov}(a, b) = -\frac{S_x}{S S_{tt}}, \quad r_{ab} = \frac{\text{Cov}(a, b)}{\sigma_a \sigma_b}$$

We thus get the values of a and b with maximum likelihood as well as the variances of a and b. Using the value of χ^2 for this a and b, we can also calculate the goodness of fit, P. In addition, the linear correlation coefficient r is related by

$$\chi^2 = (1 - r^2) \sum_{i=1}^{N} (y_i - \bar{y})^2$$

Here $$r = \frac{\displaystyle\sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_i^2}}{\sqrt{\displaystyle\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{\sigma_i^2}} \sqrt{\displaystyle\sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{\sigma_i^2}}}$$

Values of r near 1 indicate a positive correlation; r near −1 means a negative correlation and r near zero means no correlation.

The form of the equations here is given to provide good accuracy when many terms are used and to provide the variances of the parameters. Another form of the equations for a and b is simpler, but is sometimes inaccurate unless many significant digits are kept in the calculations. The minimization of χ^2 when σ_i is the same for all i gives the following equations for a and b.

$$aN + b \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i$$

$$a \sum_{i=1}^{N} x_i + b \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i x_i$$

The solution is $$b = \frac{N \displaystyle\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N \displaystyle\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2}$$

$$\bar{y} = \sum_{i=1}^{N} y_i \Big/ N, \quad \bar{x} = \sum_{i=1}^{N} x_i \Big/ N$$

$$a = \bar{y} - b\bar{x}$$

The value of χ^2 can be calculated from the formula

$$\chi^2 = \sum_{i=1}^{N} y_i^2 - a \sum_{i=1}^{N} y_i - b \sum_{i=1}^{N} y_i x_i$$

It is usually advisable to plot the observed pairs of y_i versus x_i to support the linearity assumption and to detect potential outliers. Suspected outliers can be omitted from the least-squares "fit" and then subsequently tested on the basis of the least-squares fit.

Example

Application. Brenner (*Magnetic Method for Measuring the Thickness of Non-magnetic Coatings on Iron and Steel,* National Bureau of Standards, RP1081, March 1938) suggests an alternative way of measuring the thickness of nonmagnetic coatings of galvanized zinc on iron and steel. This procedure is based on a nondestructive magnetic method as a substitute for the standard destructive stripping method. A random sample of 11 pieces was selected and measured by both methods.

Nomenclature. The calibration between the magnetic and the stripping methods can be determined through the model

$$y = a + bx + \varepsilon$$

where x = strip-method determination
y = magnetic-method determination

Sample data

Thickness, 10^{-5} In

Stripping method, x	Magnetic method, y
104	85
114	115
116	105
129	127
132	120
139	121
174	155
312	250
338	310
465	443
720	630

Computations. The normal equations are defined by

$$na + (\sum x)b = \sum y$$
$$(\sum x)a + (\sum x^2)b = \sum xy$$

For the sample

$$11a + 2743b = 2461$$
$$2743a + 1{,}067{,}143b = 952{,}517$$

with $\sum y^2 = 852{,}419$.

The solution to the normal equations is given by

$$a = 3.19960 \qquad b = .884362$$

The error sum of squares can be computed from the formula

$$\chi^2 = \sum y^2 - a \sum y - b \sum xy$$

if a sufficient number of significant digits is retained (usually six or seven digits are sufficient). Here

$$\chi^2 = 2175.14$$

If the normalized method is used in addition, the value of S_{tt} is $3.8314 \times 10^5/\sigma^2$, where σ^2 is the variance of the measurement of y. The values of a and b are, of course, the same. The variances of a and b are $\sigma_a^2 = 0.2532\sigma^2$, $\sigma_b^2 = 2.610 \times 10^{-6}\sigma^2$. The correlation coefficient is 0.996390, which indicates that there is a positive correlation between x and y. The small value of the variance for b indicates that this parameter is determined very well by the data. The residuals show no particular pattern, and the predictions are plotted along with the data in Fig. 3-58. If the variance of the measurements of y is known through repeated measurements, then the variance of the parameters can be made absolute.

Multiple Regression A general linear model is one expressed as

$$y(x) = \sum_{k=1}^{M} a_k X_k(x)$$

where the parameters are $\{a_k\}$, and the expression is linear with respect to them, and $X_k(x)$ can be any (nonlinear) functions of x, not depending on the parameters $\{a_k\}$. Then:

$$\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - \sum_{j=1}^{M} a_j X_j(x_i) \right] X_k(x_i) = 0, \quad k = 1, \ldots, M$$

This is rewritten as

$$\sum_{j=1}^{M} \left[\sum_{i=1}^{N} \frac{1}{\sigma_i^2} X_j(x_i) X_k(x_i) \right] a_j = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} X_k(x_i)$$

or as $$\sum_{j=1}^{M} \alpha_{kj} a_j = \beta_k$$

Solving this set of equations gives the parameters $\{a_j\}$, which maximize the likelihood. The variance of a_j is

$$\sigma^2(a_j) = C_{jj}$$

where $C_{jk} = \alpha_{jk}^{-1}$, or C is the inverse of α. The covariance of a_j and a_k

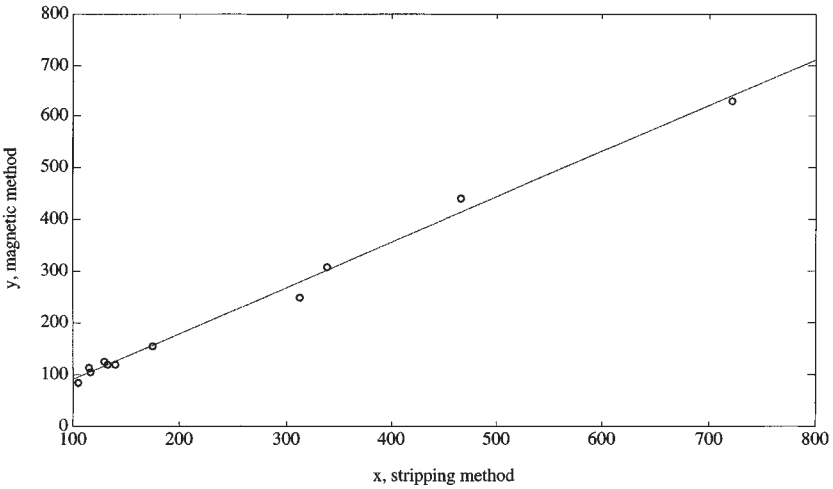

FIG. 3-58 Plot of data and correlating line.

is given by C_{jk}. If rounding errors affect the result, then we try to make the functions orthogonal. For example, using

$$X_k(x) = x^{k-1}$$

will cause rounding errors for a smaller M than

$$X_k(x) = P_{k-1}(x)$$

where P_{k-1} are orthogonal polynomials. If necessary, a singular value decomposition can be used.

Various global and piecewise polynomials can be used to fit the data. Most approximations are to be used with $M < N$. One can sometimes use more and more terms, and calculating the value of χ^2 for each solution. Then stop increasing M when the value of χ^2 no longer increases with increasing M.

Example

Application. Merriman ("The Method of Least Squares Applied to a Hydraulic Problem," *J. Franklin Inst.*, 233–241, October 1877) reported on a study of stream velocity as a function of relative depth of the stream.

Sample data

Depth*	Velocity, y, ft/s
0	3.1950
.1	3.2299
.2	3.2532
.3	3.2611
.4	3.2516
.5	3.2282
.6	3.1807
.7	3.1266
.8	3.0594
.9	2.9759

*As a fraction of total depth.

Model. Owing to the curvature of velocity with depth, a quadratic model was specified:

$$\text{Velocity} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where $x_2 = x_1^2$.

Normal equations. The three normal equations are defined by

$$(n)\hat{\beta}_0 + (\sum x_1)\hat{\beta}_1 + (\sum x_2)\hat{\beta}_2 = \sum y$$
$$(\sum x_1)\hat{\beta}_0 + (\sum x_1^2)\hat{\beta}_1 + (\sum x_1 x_2)\hat{\beta}_2 = \sum x_1 y$$
$$(\sum x_2)\hat{\beta}_0 + (\sum x_1 x_2)\hat{\beta}_1 + (\sum x_2^2)\hat{\beta}_2 = \sum x_2 y$$

For the sample data, the normal equations are

$$10\hat{\beta}_0 + 4.5\hat{\beta}_1 + 2.85\hat{\beta}_2 = 31.7616$$
$$4.5\hat{\beta}_0 + 2.85\hat{\beta}_1 + 2.025\hat{\beta}_2 = 14.08957$$
$$2.85\hat{\beta}_0 + 2.025\hat{\beta}_1 + 1.5333\hat{\beta}_2 = 8.828813$$

The algebraic solution to the simultaneous equations is

$$\hat{\beta}_0 = 3.19513 \qquad \hat{\beta}_1 = .4425 \qquad \hat{\beta}_2 = -.7653$$

The inverse of the product matrix

$$\alpha = \begin{pmatrix} 10 & 4.5 & 2.85 \\ 4.5 & 2.85 & 2.025 \\ 2.85 & 2.025 & 1.5333 \end{pmatrix}$$

is

$$\alpha^{-1} = \begin{pmatrix} .6182 & -2.5909 & 2.2727 \\ -2.5909 & 16.5530 & -17.0455 \\ 2.2727 & -17.0455 & 18.9394 \end{pmatrix}$$

The variances are then the diagonal elements of the inverse of matrix α (0.6182, 16.5530, 18.9394) times the variance of the measurement of y, σ_y^2. The value of χ^2 is 5.751×10^{-5}, the correlation coefficient $r = 0.99964$, and $\sigma = 0.002398$.

t values. A sample t value can be computed for each regression coefficient j through the identity $t_j = \hat{\beta}_j/(\hat{\sigma}\sqrt{c_{jj}})$, where c_{jj} is the (j,j) element in the inverse. For the two variables x_1 and x_2,

Coefficient	c_{jj}	Sample t value
.4425	16.55	45.3
–.7653	18.94	–73.3

Computational note. From a computational standpoint, it is usually advisable to define the variables in deviation units. For example, in the problem presented, let

$$x_1 = \text{depth} - \overline{\text{depth}}$$
$$= \text{depth} - .45$$

For expansion terms such as a square, define

$$x_2 = x_1^2 - \overline{x_1^2} \qquad (\overline{x_1^2} = .0825)$$

For the previous sample data,

Deviation units			
x_1	x_2	x_1	x_2
–.45	.12	.05	–.08
–.35	.04	.15	–.06
–.25	–.02	.25	–.02
–.25	–.02	.35	.04
–.15	–.06	.45	.12
–.05	–.08		

The resultant analysis-of-variance tables will remain exactly the same. However,

the corresponding coefficient t value for the linear coefficient will usually be improved. This is an idiosyncrasy of regression modeling. With the coded data presented, the least-squares solution is given by

$$\hat{Y} = 3.17616 - .2462x_1 - .7653x_2$$

with a corresponding t value for $\hat{\beta}_1 = -.2462$ of $t = -63.63$.

When expansion terms are used but not expanded about the mean, the corresponding t values for the generating terms should not be used. For example, if $x_3 = x_1x_2$ is used rather than the correct expansion $(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$, then the corresponding t values for x_1 and x_2 should not be used.

Nonlinear Least Squares There are no analytic methods for determining the most appropriate model for a particular set of data. In many cases, however, the engineer has some basis for a model. If the parameters occur in a nonlinear fashion, then the analysis becomes more difficult. For example, in relating the temperature to the elapsed time of a fluid cooling in the atmosphere, a model that has an asymptotic property would be the appropriate model (temp = a + b exp(−c time), where a represents the asymptotic temperature corresponding to $t \to \infty$. In this case, the parameter c appears nonlinearly. The usual practice is to concentrate on model development and computation rather than on statistical aspects. In general, nonlinear regression should be applied only to problems in which there is a well-defined, clear association between the two variables; therefore, a test of hypothesis on the significance of the fit would be somewhat ludicrous. In addition, the generalization of the theory for the associate confidence intervals for nonlinear coefficients is not well developed.

The Levenberg-Marquardt method is used when the parameters of the model appear nonlinearly (Ref. 231). We still define

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{N} \left[\frac{y_i - y(x_i; \mathbf{a})}{\sigma_i^2} \right]^2$$

and near the optimum represent χ^2 by

$$\chi^2(\mathbf{a}) = \chi_0^2 - \mathbf{d}^T \cdot \mathbf{a} + \frac{1}{2} \mathbf{a}^T \cdot \mathbf{D} \cdot \mathbf{a}$$

where $\mathbf{d}$ is an $M \times 1$ vector and $\mathbf{D}$ is an $M \times M$ matrix. We then calculate iteratively

$$\mathbf{D} \cdot (\mathbf{a}^{k+1} - \mathbf{a}^k) = -\nabla\chi^2(\mathbf{a}^k) \tag{3-89}$$

The notation a_l^k means the lth component of $\mathbf{a}$ evaluated on the kth iteration. If $\mathbf{a}^k$ is a poor approximation to the optimum, we might use steepest descent instead.

$$\mathbf{a}^{k+1} - \mathbf{a}^k = -\text{constant} \times \nabla\chi^2(\mathbf{a}^k) \tag{3-90}$$

and choose the constant somehow to decrease χ^2 as much as possible. The gradient of χ^2 is

$$\frac{\partial\chi^2}{\partial a_k} = -2 \sum_{i=1}^{N} \frac{y_i - y(x_i; \mathbf{a})}{\sigma_i^2} \frac{\partial y(x_i; \mathbf{a})}{\partial a_k} \qquad k = 1, 2, \ldots, M$$

The second derivative (in $\mathbf{D}$) is

$$\frac{\partial^2\chi^2}{\partial a_k \partial a_l} = 2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left\{ \frac{\partial y(x_i; \mathbf{a})}{\partial a_k} \frac{\partial y(x_i; \mathbf{a})}{\partial a_l} - [y_i - y(x_i; \mathbf{a})] \frac{\partial^2 y(x_i; \mathbf{a})}{\partial a_k \partial a_l} \right\}$$

Both Eq. (3-89) and Eq. (3-90) are included if we write

$$\sum_{l=1}^{M} \alpha'_{kl} (a_l^{k+1} - a_l^k) = \beta_k \tag{3-91}$$

where

$$\alpha'_{kl} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial y(x_i; \mathbf{a})}{\partial a_k} \frac{\partial y(x_i; \mathbf{a})}{\partial a_l} \qquad k \neq 1$$

$$\alpha'_{kk} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial y(x_i; \mathbf{a})}{\partial a_k} \right]^2 (1 + \lambda)$$

$$\beta_k = \sum_{i=1}^{N} \frac{y_i - y(x_i; \mathbf{a})}{\sigma_i^2} \frac{\partial y(x_i; \mathbf{a})}{\partial a_k}$$

The second term in the second derivative is dropped because it is usually small [remember that y_i will be close to $y(x_i, \mathbf{a})$]. The Levenberg-Marquardt method then iterates as follows

1. Choose $\mathbf{a}$ and calculate $\chi^2(\mathbf{a})$.
2. Choose λ, say $\lambda = 0.001$.
3. Solve Eq. (3-91) for $\mathbf{a}^{k+1}$ and evaluate $\chi^2(\mathbf{a}^{k+1})$.
4. If $\chi^2(a^{k+1}) \geq \chi^2(a^k)$ then increase λ by a factor of, say, 10 and go back to step 3. This makes the step more like a steepest descent.
5. If $\chi^2(a^{k+1}) < \chi^2(a^k)$ then update $\mathbf{a}$, i.e., use $\underline{\mathbf{a}} = a^{k+1}$, decrease λ by a factor of 10, and go back to step 3.
6. Stop the iteration when the *decrease* in χ^2 from one step to another is not statistically meaningful, i.e., less than 0.1 or 0.01 or 0.001.
7. Set $\lambda = 0$ and compute the estimated covariance matrix: $\mathbf{C} = \alpha^{-1}$. This gives the standard errors in the fitted parameters $\mathbf{a}$.

For normally distributed errors the parameter region in which χ^2 = constant can give boundaries of the confidence limits. The value of $\mathbf{a}$ obtained in the Marquardt method gives the minimum $\chi^2_{\min}$. If we set $\chi^2 = \chi^2_{\min} + \Delta\chi$ for some $\Delta\chi$ and then look at contours in parameter space where χ_1^2 = constant then we have confidence boundaries at the probability associated with χ_1^2. For example, in a chemical reactor with radial dispersion the heat transfer coefficient and radial effective heat conductivity are closely connected: decreasing one and increasing the other can still give a good fit. Thus, the confidence boundaries may look something like Fig. 3-59. The ellipse defined by $\Delta\chi^2 = 2.3$ contains 68.3 percent of the normally distributed data. The curve defined by $\Delta\chi^2 = 6.17$ contains 95.4 percent of the data.

Example

Application. Data were collected on the cooling of water in the atmosphere as a function of time.

Sample data

Time x	Temperature y
0	92.0
1	85.5
2	79.5
3	74.5
5	67.0
7	60.5
10	53.5
15	45.0
20	39.5

Model form. On the basis of the nature of the data, an exponential model was selected initially to represent the trend $y = a + be^{cx}$. In this example, the resultant temperature would approach as an asymptotic (a with c negative) the wet-bulb temperature of the surrounding atmosphere. Unfortunately, this temperature was not reported.

Using a computer package in MATLAB gives the following results: $a = 33.54$, $b = 57.89$, $c = 0.11$. The value of χ^2 is 1.83. An alternative form of model is $y = a + b/(c + x)$. For this model the results were $a = 9.872$, $b = 925.7$, $c = 11.27$, and the value of χ^2 is 0.19. Since this model had a smaller value of χ^2, it might be the chosen one, but it is only a fit of the specified data and may not be generalized beyond that. Both curve fits give an equivalent plot. The second form is shown in Fig. 3-60.

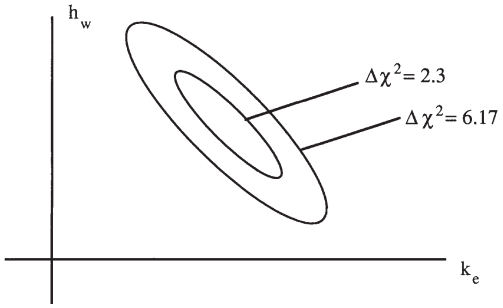

FIG. 3-59 Parameter estimation for heat transfer.

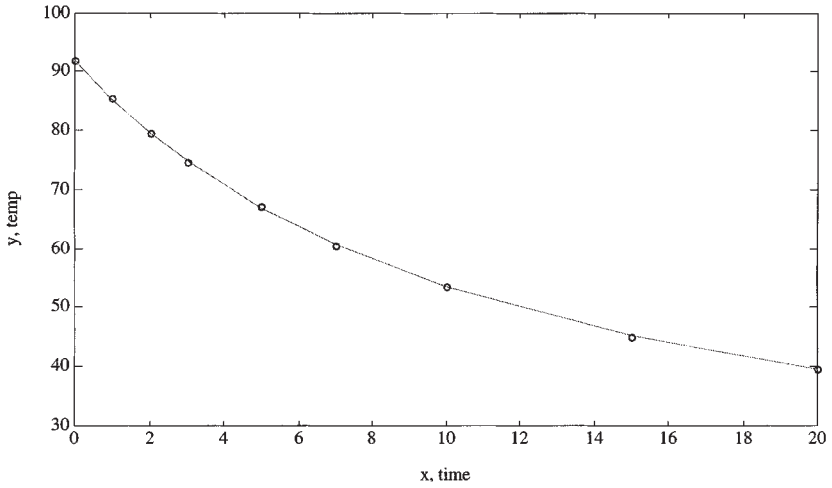

FIG. 3-60 Data in nonlinear regression example.

ERROR ANALYSIS OF EXPERIMENTS

Consider the problem of assessing the accuracy of a series of measurements. If measurements are for independent, identically distributed observations, then the errors are independent and uncorrelated. Then $\bar{y}$, the experimentally determined mean, varies about $E(y)$, the true mean, with variance σ^2/n, where n is the number of observations in $\bar{y}$. Thus, if one measures something several times today, and each day, and the measurements have the same distribution, then the variance of the means decreases with the number of samples in each day's measurement, n. Of course, other factors (weather, weekends) may make the observations on different days *not* distributed identically.

Consider next the problem of estimating the error in a variable that cannot be measured directly but must be calculated based on results of other measurements. Suppose the computed value Y is a linear combination of the measured variables $\{y_i\}$, $Y = \alpha_1 y_1 + \alpha_2 y_2 + \cdots$. Let the random variables $y_1, y_2, \ldots$ have means $E(y_1), E(y_2), \ldots$ and variances $\sigma^2(y_1), \sigma^2(y_2), \ldots$. The variable Y has mean

$$E(Y) = \alpha_1 E(y_1) + \alpha_2 E(y_2) + \cdots$$

and variance (Ref. 82)

$$\sigma^2(Y) = \sum_{i=1}^{n} \alpha_i^2 \sigma^2(y_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \alpha_i \alpha_j \operatorname{Cov}(y_i, y_j)$$

If the variables are uncorrelated and have the same variance, then

$$\sigma^2(Y) = \left(\sum_{i=1}^{n} \alpha_i^2 \right) \sigma^2$$

Next suppose the model relating Y to $\{y_i\}$ is nonlinear, but the errors are small and independent of one another. Then a change in Y is related to changes in y_i by

$$dY = \frac{\partial Y}{\partial y_1} dy_1 + \frac{\partial Y}{\partial y_2} dy_2 + \cdots$$

If the changes are indeed small, then the partial derivatives are constant among all the samples. Then the expected value of the change, $E(dY)$, is zero. The variances are given by the following equation (Refs. 25 and 40):

$$\sigma^2(dY) = \sum_{i=1}^{N} \left(\frac{\partial Y}{\partial y_i} \right)^2 \sigma_i^2$$

Thus, the variance of the desired quantity Y can be found. This gives an independent estimate of the errors in measuring the quantity Y from the errors in measuring each variable it depends upon.

Example Suppose one wants to measure the thermal conductivity of a solid (k). To do this, one needs to measure the heat flux (q), the thickness of the sample (d), and the temperature difference across the sample (ΔT). Each measurement has some error. The heat flux (q) may be the rate of electrical heat input ($\dot{Q}$) divided by the area (A), and both quantities are measured to some tolerance. The thickness of the sample is measured with some accuracy, and the temperatures are probably measured with a thermocouple to some accuracy. These measurements are combined, however, to obtain the thermal conductivity, and it is desired to know the error in the thermal conductivity. The formula is

$$k = \frac{d}{A \Delta T} \dot{Q}$$

The variance in the thermal conductivity is then

$$\sigma_k^2 = \left(\frac{k}{d} \right)^2 \sigma_d^2 + \left(\frac{k}{\dot{Q}} \right)^2 \sigma_{\dot{Q}}^2 + \left(\frac{k}{A} \right)^2 \sigma_A^2 + \left(\frac{k}{\Delta T} \right)^2 \sigma_{\Delta T}^2$$

FACTORIAL DESIGN OF EXPERIMENTS AND ANALYSIS OF VARIANCE

Statistically designed experiments consider, of course, the effect of primary variables, but they also consider the effect of extraneous variables and the interactions between variables, and they include a measure of the random error. Primary variables are those whose effect you wish to determine. These variables can be quantitative or qualitative. The quantitative variables are ones you may fit to a model in order to determine the model parameters (see the section "Least Squares"). Qualitative variables are ones you wish to know the effect of, but you do not try to quantify that effect other than to assign possible errors or magnitudes. Qualitative variables can be further subdivided into Type I variables, whose effect you wish to determine directly, and Type II variables, which contribute to the performance variability and whose effect you wish to average out. For example, if you are studying the effect of several catalysts on yield in a chemical reactor, each different type of catalyst would be a Type I variable because you would like to know the effect of each. However, each time the catalyst is prepared, the results are slightly different due to random variations; thus, you may have several batches of what purports to be the same catalyst. The variability between batches is a Type II variable. Since the ultimate use will require using different batches, you would like to know the overall effect including that variation, since knowing precisely the results from one batch of one catalyst might not be representative of the results obtained from all batches of the same catalyst. A randomized block design, incomplete block design, or Latin square design (Ref. 40), for example, all keep the effect of experimental error in the blocked variables from influencing the effect of the primary variables. Other uncontrolled variables are accounted for by introducing randomization in parts of the experimental design. To study all variables and their interaction requires a factorial design, involving all possible

combinations of each variable, or a fractional factorial design, involving only a selected set. Statistical techniques are then used to determine which are the important variables, what are the important interactions, and what the error is in estimating these effects. The discussion here is only a brief overview of the excellent Ref. 40.

Suppose we have two methods of preparing some product and we wish to see which treatment is best. When there are only two treatments, then the sampling analysis discussed in the section "Two-Population Test of Hypothesis for Means" can be used to deduce if the means of the two treatments differ significantly. When there are more treatments, the analysis is more detailed. Suppose the experimental results are arranged as shown in the table: several measurements for each treatment. The goal is to see if the treatments differ significantly from each other; that is, whether their means are different when the samples have the same variance. The hypothesis is that the treatments are all the same, and the null hypothesis is that they are different. The statistical validity of the hypothesis is determined by an analysis of variance.

Estimating the Effect of Four Treatments

	Treatment			
	1	2	3	4
	—	—	—	—
	—	—	—	—
	—	—	—	—
		—	—	—
			—	—
				—
Treatment average	—	—	—	—
Grand average		—		

The data for $k = 4$ treatments is arranged in the table. For each treatment, there are n_t experiments and the outcome of the ith experiment with treatment t is called y_{ti}. Compute the treatment average

$$\bar{y}_t = \frac{\sum_{i=1}^{n_t} y_{ti}}{n_t}$$

Also compute the grand average

$$\bar{y} = \frac{\sum_{t=1}^{k} n_t \bar{y}_t}{N}, \quad N = \sum_{t=1}^{k} n_t$$

Next compute the sum of squares of deviations from the average within the tth treatment

$$S_t = \sum_{i=1}^{n_t} (y_{ti} - \bar{y}_t)^2$$

Since each treatment has n_t experiments, the number of degrees of freedom is $n_t - 1$. Then the sample variances are

$$s_t^2 = \frac{S_t}{n_t - 1}$$

The within-treatment sum of squares is

$$S_R = \sum_{t=1}^{k} S_t$$

and the within-treatment sample variance is

$$s_R^2 = \frac{S_R}{N - k}$$

Now, if there is no difference between treatments, a second estimate of σ^2 could be obtained by calculating the variation of the treatment averages about the grand average. Thus compute the between-treatment mean square

$$s_T^2 = \frac{S_T}{k - 1}, \quad S_T = \sum_{t=1}^{k} n_t (\bar{y}_t - \bar{y})^2$$

Basically the test for whether the hypothesis is true or not hinges on a comparison of the within-treatment estimate s_R^2 (with $\nu_R = N - k$ degrees of freedom) with the between-treatment estimate s_T^2 (with $\nu_T = k - 1$ degrees of freedom). The test is made based on the F distribution for ν_R and ν_T degrees of freedom (Table 3-7).

Next consider the case that uses randomized blocking to eliminate the effect of some variable whose effect is of no interest, such as the batch-to-batch variation of the catalysts in the chemical reactor example. Suppose there are k treatments and n experiments in each treatment. The results from nk experiments can be arranged as shown in the block design table; within each block, the various treatments are applied in a random order. Compute the block average, the treatment average, as well as the grand average as before.

Block Design with Four Treatments and Five Blocks

Treatment	1	2	3	4	Block average
Block 1	—	—	—	—	—
Block 2	—	—	—	—	—
Block 3	—	—	—	—	—
Block 4	—	—	—	—	—
Block 5	—	—	—	—	—

The following quantities are needed for the analysis of variance table.

Name	Formula	dof
average	$S_A = nk\bar{y}^2$	1
blocks	$S_B = k \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2$	$n - 1$
treatments	$S_T = n \sum_{t=1}^{k} (\bar{y}_t - \bar{y})^2$	$k - 1$
residuals	$S_R = \sum_{t=1}^{k} \sum_{i=1}^{n} (y_{ti} - \bar{y}_i - \bar{y}_t + \bar{y})^2$	$(n - 1)(k - 1)$
total	$S = \sum_{t=1}^{k} \sum_{i=1}^{n} y_{ti}^2$	$N = nk$

The key test is again a statistical one, based on the value of

$$\frac{s_T^2}{s_R^2}, \quad s_T^2 = \frac{S_T}{k - 1}, \quad s_R^2 = \frac{S_R}{(n - 1)(k - 1)}$$

and the F distribution for ν_R and ν_T degrees of freedom (Table 3-7). The assumption behind the analysis is that the variations are linear (Ref. 40). There are ways to test this assumption as well as transformations to make if it is not true. Reference 40 also gives an excellent example of how the observations are broken down into a grand average, a block deviation, a treatment deviation, and a residual. For two-way factorial design in which the second variable is a real one rather than one you would like to block out, see Ref. 40.

To measure the effects of variables on a single outcome a factorial design is appropriate. In a two-level factorial design, each variable is considered at two levels only, a high and low value, often designated as a + and −. The two-level factorial design is useful for indicating trends, showing interactions, and it is also the basis for a fractional factorial design. As an example, consider a 2^3 factorial design with 3 variables and 2 levels for each. The experiments are indicated in the factorial design table.

Two-Level Factorial Design with Three Variables

	Variable		
Run	1	2	3
1	−	−	−
2	+	−	−
3	−	+	−
4	+	+	−
5	−	−	+
6	+	−	+
7	−	+	+
8	+	+	+

The main effects are calculated by calculating the difference between results from all high values of a variable and all low values of a variable; the result is divided by the number of experiments at each level. For example, for the first variable:

$$\text{Effect of variable 1} = \frac{[(y_2 + y_4 + y_6 + y_8) - (y_1 + y_3 + y_5 + y_7)]}{4}$$

Note that all observations are being used to supply information on each of the main effects and each effect is determined with the precision of a fourfold replicated difference. The advantage of a one-at-a-time experiment is the gain in precision if the variables are additive and the measure of nonadditivity if it occurs (Ref. 40).

Interaction effects between variables 1 and 2 are obtained by calculating the difference between the results obtained with the high and low value of 1 at the low value of 2 compared with the results obtained with the high and low value 1 at the high value of 2. The 12-interaction is

$$\text{12-interaction} = \frac{[(y_4 - y_3 + y_8 - y_7) - (y_2 - y_1 + y_6 - y_5)]}{2}$$

The key step is to determine the errors associated with the effect of each variable and each interaction so that the significance can be determined. Thus, standard errors need to be assigned. This can be done by repeating the experiments, but it can also be done by using higher-order interactions (such as 123 interactions in a 2^4 factorial design). These are assumed negligible in their effect on the mean but can be used to estimate the standard error (see Ref. 40). Then, calculated effects that are large compared with the standard error are considered important, while those that are small compared with the standard error are considered to be due to random variations and are unimportant.

In a fractional factorial design one does only part of the possible experiments. When there are k variables, a factorial design requires 2^k experiments. When k is large, the number of experiments can be large; for $k = 5$, $2^5 = 32$. For a k this large, Box et al. (Ref. 82, p. 376) do a fractional factorial design. In the fractional factorial design with $k = 5$, only 16 experiments are done. Cropley (Ref. 82) gives an example of how to combine heuristics and statistical arguments in application to kinetics mechanisms in chemical engineering.

DIMENSIONAL ANALYSIS

Dimensional analysis allows the engineer to reduce the number of variables that must be considered to model experiments or correlate data. Consider a simple example in which two variables F_1 and F_2 have the units of force and two additional variables L_1 and L_2 have the units of length. Rather than having to deduce the relation of one variable on the other three, $F_1 = \text{fn}\,(F_2, L_1, L_2)$, dimensional analysis can be used to show that the relation must be of the form $F_1/F_2 = \text{fn}\,(L_1/L_2)$. Thus considerable experimentation is saved. Historically, dimensional analysis can be done using the Rayleigh method or the Buckingham pi method. This brief discussion is equivalent to the Buckingham pi method but uses concepts from linear algebra; see Ref. 13 for further information.

The general problem is posed as finding the minimum number of variables necessary to define the relationship between n variables. Let $\{Q_i\}$ represent a set of fundamental units, like length, time, force, and so on. Let $[P_i]$ represent the dimensions of a physical quantity P_i; there are n physical quantities. Then form the matrix α_{ij}

	$[P_1]$	$[P_2]$	⋯	$[P_n]$
Q_1	α_{11}	α_{12}	⋯	α_{1n}
Q_2	α_{21}	α_{22}	⋯	α_{2n}
⋯				
Q_m	α_{m1}	α_{m2}	⋯	α_{mn}

in which the entries are the number of times each fundamental unit appears in the dimensions $[P_i]$. The dimensions can then be expressed as follows.

$$[P_i] = Q_1^{\alpha_{1i}} Q_2^{\alpha_{2i}} \cdots Q_m^{\alpha_{mi}}$$

Let m be the rank of the α matrix. Then $p = n - m$ is the number of dimensionless groups that can be formed. One can choose m variables $\{P_i\}$ to be the basis and express the other p variables in terms of them, giving p dimensionless quantities.

Example: Buckingham Pi Method—Heat-Transfer Film Coefficient It is desired to determine a complete set of dimensionless groups with which to correlate experimental data on the film coefficient of heat transfer between the walls of a straight conduit with circular cross section and a fluid flowing in that conduit. The variables and the dimensional constant believed to be involved and their dimensions in the engineering system are given below:

Film coefficient = $h = (F/L\theta T)$
Conduit internal diameter = $D = (L)$
Fluid linear velocity = $V = (L/\theta)$
Fluid density = $\rho = (M/L^3)$
Fluid absolute viscosity = $\mu = (M/L\theta)$
Fluid thermal conductivity = $k = (F/\theta T)$
Fluid specific heat = $c_p = (FL/MT)$
Dimensional constant = $g_c = (ML/F\theta^2)$

The matrix α in this case is as follows.

		$[P_i]$							
		h	D	V	ρ	μ	k	C_p	g_c
Q_j	F	1	0	0	0	0	1	1	−1
	M	0	0	0	1	1	0	−1	1
	L	−1	1	1	−3	−1	0	1	1
	θ	−1	0	−1	0	−1	−1	0	−2
	T	−1	0	0	0	0	−1	−1	0

Here $m \le 5$, $n = 8$, $p \ge 3$. Choose D, V, μ, k, and g_c as the primary variables. By examining the 5×5 matrix associated with those variables, we can see that its determinant is not zero, so the rank of the matrix is $m = 5$; thus, $p = 3$. These variables are thus a possible basis set. The dimensions of the other three variables h, ρ, and C_p must be defined in terms of the primary variables. This can be done by inspection, although linear algebra can be used, too.

$$[h] = D^{-1}k^{+1}; \text{ thus } \frac{h}{D^{-1}k} = \frac{hD}{k} \text{ is a dimensionless group}$$

$$[\rho] = \mu^{1}V^{-1}D^{-1}; \text{ thus } \frac{\rho}{\mu^{1}V^{-1}D^{-1}} = \frac{\rho VD}{\mu} \text{ is a dimensionless group}$$

$$[C_p] = k^{+1}\mu^{-1}; \text{ thus } \frac{C_p}{k^{+1}\mu^{-1}} = \frac{C_p\mu}{k} \text{ is a dimensionless group}$$

Thus, the dimensionless groups are

$$\frac{[P_i]}{Q_1^{\alpha_{1i}} Q_2^{\alpha_{2i}} \cdots Q_m^{\alpha_{mi}}} : \frac{hD}{k}, \frac{\rho VD}{\mu}, \frac{C_p\mu}{k}$$

The dimensionless group hD/k is called the Nusselt number, N_{Nu}, and the group $C_p\mu/k$ is the Prandtl number, N_{Pr}. The group $DV\rho/\mu$ is the familiar Reynolds number, N_{Re}, encountered in fluid-friction problems. These three dimensionless groups are frequently used in heat-transfer-film-coefficient correlations. Functionally, their relation may be expressed as

$$\phi(N_{\text{Nu}}, N_{\text{Pr}}, N_{\text{Re}}) = 0 \qquad (3\text{-}91)$$

or as

$$N_{\text{Nu}} = \phi_1(N_{\text{Pr}}, N_{\text{Re}})$$

It has been found that these dimensionless groups may be correlated well by an equation of the type

$$hD/k = K(c_p\mu/k)^a(DV\rho/\mu)^b$$

in which K, a, and b are experimentally determined dimensionless constants. However, any other type of algebraic expression or perhaps simply a graphical relation among these three groups that accurately fits the experimental data would be an equally valid manner of expressing Eq. (3-91).

Naturally, other dimensionless groups might have been obtained in the example by employing a different set of five repeating quantities that would not form a dimensionless group among themselves. Some of these groups may be found among those presented in Table 3-8. Such a complete set of three dimensionless groups might consist of Stanton, Reynolds, and Prandtl numbers or of Stanton, Peclet, and Prandtl numbers. Also, such a complete set different from that obtained in the preceding example will result from a multiplication of appropriate powers of the Nusselt, Prandtl, and Reynolds numbers. For such a set to be complete, however, it must satisfy the condition that each of the three dimensionless groups be independent of the other two.

TABLE 3-8 Dimensionless Groups in the Engineering System of Dimensions

Biot number	N_{Bi}	hL/k
Condensation number	N_{Co}	$(h/k)(\mu^2/\rho^2 g)^{1/3}$
Number used in condensation of vapors	N_{Cv}	$L^3\rho^2 g\lambda/k\mu\Delta t$
Euler number	N_{Eu}	$g_c(-dp)/\rho V^2$
Fourier number	N_{Fo}	$k\theta/\rho c L^2$
Froude number	N_{Fr}	V^2/Lg
Graetz number	N_{Gz}	wc/kL
Grashof number	N_{Gr}	$L^3\rho^2\beta g\Delta t/\mu^2$
Mach number	N_{Ma}	V/V_a
Nusselt number	N_{Nu}	hD/k
Peclet number	N_{Pe}	$DV\rho c/k$
Prandtl number	N_{Pr}	$c\mu/k$
Reynolds number	N_{Re}	$DV\rho/\mu$
Schmidt number	N_{Sc}	$\mu/\rho D_v$
Stanton number	N_{St}	$h/cV\rho$
Weber number	N_{We}	$LV^2\rho/\sigma g_c$

PROCESS SIMULATION

Classification Process simulation refers to the activity in which mathematical models of chemical processes and refineries are modeled with equations, usually on the computer. The usual distinction must be made between steady-state models and transient models, following the ideas presented in the introduction to this section. In a chemical process, of course, the process is nearly always in a transient mode, at some level of precision, but when the time-dependent fluctuations are below some value, a steady-state model can be formulated. This subsection presents briefly the ideas behind steady-state process simulation (also called flowsheeting), which are embodied in commercial codes. The transient simulations are important for designing startup of plants and are especially useful for the operating of chemical plants.

Process Modules The usual first step in process simulation is to perform a mass and energy balance for a chosen process. The most important aspect of the simulation is that the thermodynamic data of the chemicals be modeled correctly. The computer results of vapor-liquid equilibria, for example, must be checked against experimental data to insure their validity before using the data in more complicated computer calculations. At this first level of detail, it is not necessary to know the internal parameters for all the units, since what is desired is just the overall performance. For example, in a heat exchanger design, it suffices to know the heat duty, the total area, and the temperatures of the output streams; the details like the percentage baffle cut, tube layout, or baffle spacing can be specified later when the details of the proposed plant are better defined. Each unit operation is modeled by a subroutine, which is governed by equations (presented throughout this book). Some of the inputs to the units are known, some are specified by the user as design variables, and some are to be found using the simulation. It is important to know the number of degrees of freedom for each option of the unit operation, because at least that many parameters must be specified in order for the simulation to be able to calculate unit outputs. Sometimes the quantities the user would like to specify are targets, and parameters in the unit operation are to be changed to meet that target. This is not always possible, and the designer will have to adjust the parameters of the unit operation to achieve the desired target, possibly using the convergence tools discussed below. For example, in a reaction/separation system, if there is an impurity that must be purged, a common objective is to set the purge fraction so that the impurity concentration into the reactor is kept at some moderate value. Yet the solution techniques do not readily lend themselves to this connection, so convergence strategies must be employed.

Solution Strategies Consider a chemical process consisting of a series of units, such as distillation towers, reactors, and so forth. If the feed to the process is known and the operating parameters of the unit operations are specified by the user, then one can begin with the first unit, take the process input, calculate the unit output, carry that output to the input of the next unit, and continue the process. In this way, one can simulate the entire process. However, if the process involves a recycle stream, as nearly all chemical processes do, then when the calculation is begun, it is discovered that the recycle stream is unknown. Thus the calculation cannot begin. This situation leads to the need for an iterative process: the flow rates, temperature, and pressure of the unknown recycle stream are guessed and the calculations proceed as before. When one reaches the end of the process, where the recycle stream is formed to return to the inlet, it is necessary to check to see if the recycle stream is the same as assumed. If not, an iterative procedure must be used to cause convergence. The techniques like Wegstein (see "Numerical Solution of Nonlinear Equations in One Variable") can be used to accelerate the convergence. When doing these iterations, it is useful to analyze the process using precedence ordering and tearing to minimize the number of recycle loops (Refs. 201, 242, 255, and 293). When the recycle loops interact with one another the iterations may not lead to a convergent solution.

The designer usually wants to specify stream flow rates or parameters in the process, but these may not be directly accessible. For example, the desired separation may be known for a distillation tower, but the simulation program requires the specification of the number of trays. It is left up to the designer to choose the number of trays that lead to the desired separation. In the example of the purge stream/reactor impurity, a controller module may be used to adjust the purge rate to achieve the desired reactor impurity. This further complicates the iteration process.

An alternative method of solving the equations is to solve them as simultaneous equations. In that case, one can specify the design variables and the desired specifications and let the computer figure out the process parameters that will achieve those objectives. It is possible to overspecify the system or give impossible conditions. However, the biggest drawback to this method of simulation is that large sets (10,000s) of algebraic equations must be solved simultaneously. As computers become faster, this is less of an impediment.

For further information, see Refs. 90, 175, 255, and 293. For information on computer software, see the Annual CEP Software Directory (Ref. 8) and other articles (Refs. 7 and 175).

INTELLIGENT SYSTEMS IN PROCESS ENGINEERING

REFERENCES: General, 232, 248, 258, 275, 276. Knowledge-Based Systems, 49, 232, 275. Neural Networks, 54, 140. Qualitative Simulation, 178, 292. Fuzzy Logic, 94. Genetic Algorithms, 121. Applications, 15, 24, 205, 232, 250, 262, 294.

Intelligent system is a term that refers to computer-based systems that include knowledge-based systems, neural networks, fuzzy logic and fuzzy control, qualitative simulation, genetic algorithms, natural language understanding, and others. The term is often associated with a variety of computer programming languages and/or features that are used as implementation media, although this is an imprecise use. Examples include object-oriented languages, rule-based languages, prolog, and lisp. The term *intelligent system* is preferred over the term *artificial intelligence.* The three intelligent-system technologies currently seeing the greatest amount of industrial application are knowledge-based systems, fuzzy logic, and artificial neural networks. These technologies are components of distributed systems. Mathematical models, conventional numeric and statistical approaches, neural networks, knowledge-based systems, and the like, all have their place in practical implementation and allow automation of tasks not well-treated by numerical algorithms.

Fundamentally, intelligent-system techniques are modeling techniques. They allow the encoding of qualitative models that draw upon experience and expertise, thereby extending modeling capacity beyond mathematical description. An important capability of intelligent system techniques is that they can be used not only to model physical behaviors but also decision-making processes. Decision processes reflect the selection, application, and interpretation of highly relevant pieces of information to draw conclusions about complex situations. Activity-specific decision processes can be expressed at a functional level, such as diagnosis, design, planning, and scheduling, or as their generic components, such as classification, abduction, and simulation. Decision process models address how information is organized and structured and then assimilated into active decisions.

Knowledge-Based Systems Knowledge-based system (KBS) approaches capture the structural and information processing features of qualitative problem solving associated with sequential consideration, selection, and search. These technologies not only provide the means of capturing decision-making knowledge but also offer a medium for exploiting efficient strategies used by experts.

KBSs, then, are computer programs that model specific ways of organizing problem-specific fragments of knowledge and then searching through them by establishing appropriate relationships to reach correct conclusions. *Deliberation* is a general label for the algorithmic process for sorting through the knowledge fragments. The basic components of KBSs are knowledge representation (structure) and search. They are the programming mechanisms that facilitate the use and application of the problem-specific knowledge appropriate to solving the problem. Together they are used to form conclusions, decisions, or interpretations in a symbolic form. See Refs. 49, 232, and 275.

Qualitative simulation is a specific KBS model of physical processes that are not understood well enough to develop a physics-based numeric model. Corrosion, fouling, mechanical wear, equipment failure, and fatigue are not easily modeled, but decisions about them can be based on qualitative reasoning. See Refs. 178 and 292.

Qualitative description of physical behaviors require that each continuous variable space be quantized. Quantization is typically based on landmark values that are boundary points separating qualitatively distinct regions of continuous values. By using these qualitative quantity descriptions, dynamic relations between variables can be modeled as qualitative equations that represent the structure of the system. The solution to the equations represents the possible sequences of qualitative states as well as the explanations for changes in behaviors.

Building and explaining a complex model requires a unified view called an *ontology.* Methods of qualitative reasoning can be based on different viewpoints; the dominant viewpoints are device, process, and constraints. Behavior generation is handled with two approaches: (1) simulating successive states from one or more initial states, and (2) determining all possible state-to-state transitions once all possible states are determined.

Fuzzy Logic Fuzzy logic is a formalism for mapping between numerical values and qualitative or linguistic interpretations. This is useful when it is difficult to define precisely such terms as "high" and "low," since there may be no fixed threshold. Fuzzy sets use the concept of degree of membership to overcome this problem. Degree of membership allows a descriptor to be associated with a range of numeric values but in varying degrees. A fuzzy set is explicitly defined by a degree of membership for each linguistic variable that is applicable, $m_A(x)$ where m_A is the degree of membership for linguistic variable *A.* For fuzzy sets, logical operators, such as complement (NOT), intersection (AND), and union (OR) are defined. The following are typical definitions.

$$\text{NOT: } m_{\text{NOT } A}(x) = 1 - m_A(x)$$

$$\text{AND: } m_{A \text{ AND } B}(x) = \min\,[m_A(x), m_B(x)]$$

$$\text{OR: } m_{A \text{ OR } B}(x) = \max\,[m_A(x), m_B(x)]$$

Using these operators, fuzzy inference mechanisms are then developed to manipulate rules that include fuzzy values. The largest difference between fuzzy inference and ordinary inference is that fuzzy inference allows "partial match" of input and produces an "interpolated" output. This technology is useful in control also. See Ref. 94.

Artificial Neural Networks An artificial neural network (ANN) is a collection of computational units that are interconnected in a network. Knowledge is captured in the form of weights, and input-output mappings are produced by the interactions of the weights and the computational units. Each computational unit combines weighted inputs and generates an output base on an activation function. Typical activation functions are (1) specified limit, (2) sigmoid, and (3) gaussian. ANNs can be feedforward, with multiple layers of intermediate units, or feedback (sometimes called recurrent networks).

The ability to generalize on given data is one of the most important performance characteristics. With appropriate selection of training examples, an optimal network architecture, and appropriate training, the network can map a relationship between input and output that is complete but bounded by the coverage of the training data.

Applications of neural networks can be broadly classified into three categories:

1. Numeric-to-numeric transformations are used as empirical mathematical models where the adaptive characteristics of neural networks learn to map between numeric sets of input-output data. In these modeling applications, neural networks are used as an alternative to traditional data regression schemes based on regression of plant data. Backpropagation networks have been widely used for this purpose.
2. Numeric-to-symbolic transformations are used in pattern-recognition problems where the network is used to classify input data vectors into specific labeled classes. Pattern recognition problems include data interpretation, feature identification, and diagnosis.
3. Symbolic-to-symbolic transformations are used in various symbolic manipulations, including natural language processing and rule-based system implementation. See Refs. 54 and 140.

blank page 3-92