

STATISTICAL ANALYSIS

BY

EDMUND E. DAY

PROFESSOR OF ECONOMICS, AND DEAN OF THE
SCHOOL OF BUSINESS ADMINISTRATION
UNIVERSITY OF MICHIGAN

New York

THE MACMILLAN COMPANY

1925

All rights reserved

PRINTED IN THE UNITED STATES OF AMERICA

COPYRIGHT, 1925,
BY THE MACMILLAN COMPANY.

Set up and electrotyped. Published November, 1925.

Norwood Press
J. S. Cushing Co. — Berwick & Smith Co.
Norwood, Mass., U.S.A.

TO MY COUNSELLOR AND FRIEND
ALLYN A. YOUNG
WHOSE WORK IN STATISTICS HAS COMBINED IN
RARE DEGREE MASTERY OF TECHNIQUE
BREADTH OF SCHOLARSHIP AND
SOUNDNESS OF JUDGMENT

PREFACE

THIS book affords a general introduction to statistical method. It is intended primarily for those who are making their first acquaintance with the subject. The requirements of classroom use have been kept constantly in mind, and every effort has been made to make the material teachable. At the same time, the treatment of the subject is comprehensive and designed to start the intelligent student on the road to a complete grasp of general statistical method.

The emphasis of the book is upon the analysis, rather than upon the collection and tabulation, of statistical material. It is the author's opinion that the great majority of students of statistics are not directly interested in the technique of collection and tabulation. They are concerned with these phases of statistical method only in so far as they affect the subsequent analysis of data. Furthermore, the more important aspects of collection and tabulation can be fully understood only after the nature and purposes of analysis have been mastered. (True, some acquaintance with the technicalities of collection and tabulation is desirable at an early stage of the study of analysis; consequently brief accounts of these are given in a series of appendices.) In general, therefore, attention may be profitably focused at the outset on the problems of analysis, and this is the plan of the present volume.

Special care has been taken throughout the book to make clear the logical relationships of the parts. This results in rather more formalism than would be desirable in any treatment other than that of an introductory text. But it is highly important at the beginning of the study of statistics to learn to place correctly in the general scheme of statistical method the processes that are to be undertaken in any particular analysis. The student should be made to see how the individual parts are related to the whole.

In more extended form, the title of this book might be given as "The Logic of Statistical Analysis." Several years of experience in teaching statistics have led the author to believe that it is much more important

for the beginning student to learn the logical setting of statistical analysis than it is for him to gain facility in any set of technical processes. Some command of technical procedure, of course, is necessary. But, after all, the objectives of statistical work lie in the interpretation of statistical results; and interpretation can never be wisely undertaken save with full recognition of the logical implications and limitations of the processes that have gone before. At the risk, perhaps, of being somewhat meticulous, if not pedantic, the present treatment has been written with the avowed intention of making perfectly clear the *logic* of statistical procedure.

In view of this underlying purpose, it has not been thought necessary to enter upon the more refined mathematical phases of statistical method. Nothing in the book should prove inconsistent with the findings of the most advanced mathematical statistics. But no command of advanced mathematics is expected of those for whom the book has been written. For those who have special interests or preparation along mathematical lines, supplementary studies can be readily undertaken. Furthermore, it is to be hoped that some of those who secure through this book their introduction to statistics will be encouraged to pursue the subject into its fascinating and important mathematical ramifications. A serviceable start in the understanding of statistical analysis may be made, however, with a modicum of purely mathematical procedure, and in the present exposition of the subject the mathematical requirements have been reduced to a minimum.

Attention should be called to one special feature of the book: the treatment of methods of graphic representation. It is customary in texts on statistical method to devote special chapters to statistical graphics. From a pedagogical point of view, this plan has certain advantages. It involves, however, the serious danger of divorcing the practices of statistical graphics from the processes of statistical analysis. Most of the abuses of statistical graphics have occurred because diagrams have not been made to exhibit accurately the results of analysis. A closer inter-relation of analytical and graphic practices is desirable. In the present text, therefore, questions of graphic method are dealt with wherever they arise in the explanation of the manifold phases of analysis.

It is quite impossible for the author to state explicitly the extent of his indebtedness to others. With other workers in the field, he has profited, in part perhaps unconsciously, from the labors of many scholars.

However, some of those whose contributions have been found especially valuable should be specifically mentioned. All readers who are at all familiar with the literature of the subject will quickly sense the author's heavy indebtedness to such eminent authorities as Bowley, Fisher, Mitchell, Pearson, Persons, and Yule. Numerous footnotes throughout the book indicate sources which have been explicitly utilized at particular points. The author's colleagues, Professors O. W. Blackett, J. P. Mitchell, and C. S. Yoakum, have examined the manuscript and offered numerous valuable comments. Former students — among whom perhaps should be particularly mentioned Professors W. A. Berridge and M. B. Hexter — have made many helpful suggestions. Finally, the author's secretary, Miss Carolyn E. Allen, has given invaluable and tireless assistance without which the book could hardly have been published at this time, if at all.

EDMUND E. DAY

ANN ARBOR, MICHIGAN
September 20, 1925

TABLE OF CONTENTS

INTRODUCTION

CHAPTER	PAGE
I. THE SIGNIFICANCE OF STATISTICS	I
II. VARIABLES AND STATISTICAL UNITS	10
A. Variables	10
B. Statistical Units	17
III. ORIGINAL OBSERVATION	25
A. Observational Limits	25
B. Differentiation	28
C. Complete <i>vs.</i> Partial Observation	32
D. Accuracy	33
IV. CLASSIFICATION AND STATISTICAL SERIES	36
A. Classification	36
B. Statistical Series	42

FORMULATION OF STATISTICAL DISTRIBUTIONS

V. CLASSIFICATIONS NOT IN SERIAL FORM	48
A. Classifications by Kind	48
B. Classifications of Degree (Qualitative)	57
VI. FREQUENCY DISTRIBUTIONS	62
A. Simple Form	62
B. Cumulative Form	80
VII. SPATIAL DISTRIBUTIONS	90
VIII. TEMPORAL DISTRIBUTIONS	104
A. Simple Form	104
B. Cumulative Form	113

ANALYSIS OF FREQUENCY DISTRIBUTIONS

CHAPTER		PAGE
IX.	TYPES OF FREQUENCY DISTRIBUTIONS	118
X.	TYPICAL SIZE: AVERAGES	134
	A. The Definition and Computation of Averages	134
	B. The Selection and Use of Averages	153
XI.	DISPERSION	163
	A. Variability, or the Extent of Dispersion	163
	B. Asymmetry, or Skewness	175

ANALYSIS OF PAIRED VARIABLES: CORRELATION

XII.	THE MEANING OF CORRELATION	180
XIII.	THE MEASUREMENT OF CORRELATION	191
	A. The Pearsonian Coefficient	193
	B. The Line of Regression	199
	C. The Correlation Ratio	201
	D. Correlation from Rank Data	206

ANALYSIS OF SPATIAL SERIES

XIV.	SPATIAL SERIES AND THEIR GRAPHIC COMPARISON	211
	A. The Nature and Formulation of Spatial Series	211
	B. Graded Maps	215
	C. Graphic Comparison of Spatial Series	218

ANALYSIS OF TIME SERIES

XV.	THE NATURE OF TIME SERIES	231
XVI.	INCREMENTS AND RATES OF CHANGE	246
XVII.	EVOLUTIONARY MOVEMENTS: SECULAR TREND	258
XVIII.	PERIODIC MOVEMENTS: SEASONAL VARIATION	281
XIX.	RESIDUAL MOVEMENTS: CYCLICAL AND IRREGULAR FLUCTUATIONS	302
	A. Episodic Movements	302
	B. Cyclical Fluctuations	306
	C. Accidental Fluctuations	310

CONTENTS

xiii

CHAPTER	PAGE
XX. CORRESPONDENCE AND CORRELATION IN TIME SERIES ;	
LAG	313
A. The Extent of Correspondence or Correlation .	313
B. Time Relationship: Lead and Lag	322
ANALYSIS OF GROUPED VARIABLES: INDEX NUMBERS	
XXI. THE NATURE AND PURPOSE OF INDEX NUMBERS .	328
XXII. UNWEIGHTED INDEX NUMBERS	342
XXIII. WEIGHTED INDEX NUMBERS	352
CONCLUSION	
XXIV. THE NATURE OF STATISTICAL RESULTS	368
APPENDICES	
A. THE COLLECTION OF PRIMARY DATA	383
B. THE PRELIMINARY EXAMINATION OF SECONDARY DATA .	389
C. THE MECHANICS OF TABULATION	393
D. RULES FOR THE CONSTRUCTION OF STATISTICAL TABLES	403
E. RULES FOR THE CONSTRUCTION OF STATISTICAL CHARTS .	411
F. TABLE OF LOGARITHMS	421
G. TABLE OF SQUARES, SQUARE ROOTS, AND RECIPROCALs — 1 TO 1,000	426
H. RELATIONSHIP BETWEEN r AND ρ	452
I. LIST OF GENERAL REFERENCES	453
INDEX	455

TABLES

TABLE	PAGE
1. Immigrant Aliens Admitted to the United States, Fiscal Year Ended June 30, 1922	38-40
A. By Race	38
B. By Port of Entry	39
C. By Month of Entry	39
D. By Sex	40
E. By Age	40
2. A. Simple Classification of Members of College Class by Age	41
B. Multiple Classification of Members of College Class by Age and Height	41
3. Farm Price of Raw Cotton in the United States by Years, 1910-1923	43
4. Distribution of Newsboys of Cincinnati in 1918 by Age	44
5. Average Weekly Earnings of 700 Newsboys of Cincinnati in 1918 by Age	44
6. Percentage of Foreign Born in Population of New England, January 1, 1920, by States	45
7. Cotton Crop of the United States (Exclusive of Linters) by Crop-Years, 1917-1922	45
8. Average Number of Wage Earners in Major Industrial Groups of Manufacturing Establishments in the United States in 1919	49
9. Causes of Business Failures in the United States in 1922	49
10. Amounts of Different Kinds of Money in the United States in Banks Other than Federal Reserve Banks, and in Circulation, June 30, 1921	50
11. General Departmental Expenses of State Governments in the United States in 1921	50
12. Percentage Distribution of General Departmental Expenses of State Governments in the United States in 1921	52
13. Causes of Business Failures in the United States in 1922	52

TABLE	PAGE
14. Percentage of National Income Contributed by Various Industries in the United States (Average 1909 to 1918)	54
15. Distribution of Steamship Tonnage of the United States, by Speed, March 31, 1919	62
16. Original Reports by 120 Unions Giving Percentage of Members Unemployed, July 1, 1920	63
17. Distribution of 120 Reporting Trade Unions according to Percentage of Members Unemployed, July 1, 1920	64
17a. Array of Original Reports of 120 Unions Giving Percentage of Members Unemployed	64
18. Number of Garment Manufacturing Concerns in 1917, Grouped according to Invested Capital	65
19. Length-of-Service Distribution of Employees Who Left during Six Months' Period	67
20. Simple Frequency Distribution of Price Relatives of 96 Commodities Reported by Bradstreet's for May 1, 1914, and May 1, 1922 (May 1, 1914 = 100)	68
21. Frequency Distribution of Price Relatives of Table 20 on Ratio (or Logarithmic) Scale	69
22. Percentage Distribution of Weekly Wages among Male Weavers in Cotton Mills of States A and B	78
23. Frequency Distributions of Percentage of Members Unemployed July 1, 1920, among 120 Trade Unions	81
24. Cumulative Percentage Distribution of Weekly Wages among Male Weavers in Cotton Mills of State A	83
25. Data for Lorenz Curve of Wage Earners and Wage Receipts in Given Wage-Earning Group	88
26a. Distribution of Population of Kansas by Age, January 1, 1920	90
26b. Distribution of Population of Kansas by Counties, January 1, 1920	91
27. Number of Persons Living within Five Minutes' Walk of Successive Stops on Suburban Trolley Line	92
28. Monthly Steel Ingot Output in the United States January, 1920, to June, 1922	108
29. Monthly Production of Portland Cement in the United States in 1923	114
30. Cumulative Scheduled and Actual Monthly Production of Aircraft, 1918	117

TABLE	PAGE
31. Frequency Distribution of Height among Members of Freshman Class in 1913 in Ohio State University . . .	119
32. Frequency Distribution of Percentage of Freight Traffic to Total Traffic on 131 American Railroads in 1920 . . .	120
33. Frequency Distribution of Ratio of Assessed to True Value of Property in 119 Counties of Kentucky . . .	120
34. Frequency Distribution of Number of Heads per Toss in 1000 Tosses of Six Coins	121
35. Terms of the Binomial Series $10,000 (p + q)^{20}$ for Values of p from 0.5 to 0.1	124
36. Frequency Distribution of Weekly Wage-Rates among Male Weavers in Selected Cotton Mills in the United States	126
37. Frequency Distribution of Reported Income among New York State Residents in the Legal Profession, 1920 . . .	127
38. Frequency Distribution of Number of Children per Wife in Fall River, Massachusetts	128
39. Frequency Distribution of Hourly Rates of Wages among Female Cotton Spinners	130
40. Frequency Distribution of Hourly Rates of Wages among Female Cotton Spinners in Groups A and B . . .	131
41. Calculation of Arithmetic Mean from Simple Frequency Distribution: Long Method	137
42. Calculation of Arithmetic Mean from Simple Frequency Distribution: Short-Cut Method	139
43. Calculation of Geometric Mean from Simple Frequency Distribution	141
44. Calculation of Harmonic Mean from Simple Frequency Distribution	143
45. Interpolation of Median in Simple Frequency Distribution	145
46. Determination of Mode by Method of Successive Regrouping	150
47. Calculation of Weighted Average	152
48. Calculation of Average Deviation from Frequency Distribution, Using Deviations from Median	166
49. Calculation of Average Deviation from Frequency Distribution, Using Deviations from Arbitrary Origin . . .	167
50. Calculation of Standard Deviation from Frequency Distribution: Long Method	168

TABLE	PAGE
51. Calculation of Standard Deviation from Frequency Distribution: Short-Cut Method	169
52. Location of Quartiles in Frequency Distribution	171
53. Comparison of Measures and Coefficients of Dispersion	175
54. Frequency Distribution of Number in Litter among Litters of Mice	178
55. Correlative Distribution of (A) per Capita Wealth and (B) per Capita Registration of Automobiles among 46 States in 1922	183
56. Correlative Distribution of the Quantity of Output and the Size of Labor Force in Coal Mines Delivering more than 100,000 Tons during Given Year	184
57. Correlative Distribution of Age and Weekly Earnings among Cincinnati Newsboys under Sixteen Years of Age and Earning less than Three Dollars a Week	185
58. Correlative Distribution of Paired Values of X and Y (Condition of Perfect Independence)	186
59. Correlative Distribution of Paired Values of X and Y (Condition of Perfect Correlation)	187
60. Average Weekly Earnings of 624 Newsboys of Cincinnati by Age	188
61. Calculation of Correlation Coefficient from Simple Series: Long Method	195
62. Calculation of Correlation Coefficient from Simple Series: Shorter Method	197
63. Calculation of Correlation Coefficient from Correlation Table: Short-Cut Method	198
64. Correlative Distribution of X and Y , Correlation Perfect but Non-Linear	202
65. Relation between Business Produced in 1921 and Length of Service with Company among Salesmen of Insurance Company	203
66. Calculation of Correlation Coefficient and Ratio from Correlation Table: Short-Cut Method	205
67. Rank of the 48 States in 1922, in per Capita (<i>a</i>) Property Values and (<i>b</i>) Automobile Registrations	207
68. Calculation of ρ from Rank Data	208
69. Monthly Mean of Free Air Temperatures, July 1, 1907, to June 30, 1910	213

TABLES

xix

TABLE	PAGE
70A. Proportion of Ford Passenger Cars to Total Passenger-Car Registrations, January, 1923, by States	227
70B. Per cent Rural Population to Total Population, 1920, by States	227
71. Comparable Grades for Two Shaded Maps	228
72. Tons of Cargo on Commercial Vessels Passing through the Panama Canal	231
73. Raw Cotton Consumed (Exclusive of Linters), January, 1914, to December, 1920	232
74. Ratio of Coin and Bullion to the Sum of Total Deposits and Note Circulation, of the Bank of England	232
75. Population of the United States (Exclusive of Outlying Possessions), 1790-1920	236
76. Imports of Coffee into the United States by Years, 1910-19	237
77. Calls on Telephone Exchange during Month of June, 19—	240
78. Precipitation for the Month of May in Each Year 1871-1916, Nashville, Tenn.	241
79. Monthly Tonnage of Pig Iron Produced in the United States, 1903-1919	243
80. Population Growth of the United States (Exclusive of Outlying Possessions), 1790-1920	246
81. Annual Production of Cigarettes in the United States by Years, 1899-1919	251
82. Monthly Retail Sales of — Automobiles, January, 1922, to December, 1924	256
83. Sugar Melted at Atlantic Ports by Months, January, 1919, to December, 1921	262
84. Calculation of Slope and Intercept of Straight Line of Best Fit by Method of Moments: Odd Number of Items in Period of Fit	270
85. Calculation of Slope and Intercept of Straight Line of Best Fit by Method of Moments: Even Number of Items in Period of Fit	272
86. Calculation of Constants of Compound-Interest Curve of Best Fit by Method of Moments Applied to Logarithms	275
87. Elimination of Trend through Expression of Original Items as Percentages of Ordinate of Trend	277
88. Elimination of Trend through Comparison of Series of the Same Species	279

TABLE	PAGE
89. Mean Monthly Flows of Colorado, Niagara, and Tennessee Rivers	282
90. Ratio of Monthly Deaths to Average Number for Each Month	283
91. Selected Monthly Indices of Seasonal Variation	283
92. Typical Daily Fluctuations in Number of Telephone Toll Calls during June, 1923	284
93. Traffic Changes on Fifth Avenue at Forty-Second Street, New York City, between the Hours of Seven A.M. and Seven P.M.	285
94. Average Hourly Wind Velocity in Miles, San Francisco, 1891-1910	286
95. Rainfall in Inches on the Wachusett Watershed, 1897-1920	288
96. Link Relatives of Monthly Pig-Iron Production	291
97. Frequency Distributions of Link Relatives of Monthly Pig-Iron Production, 1904-1914	292
98. Logarithmic Correction of Median Link Relatives in Computation of Seasonal Index: Form A	293
99. Logarithmic Correction of Median Link Relatives in Computation of Seasonal Index: Form B	295
100. Arithmetic Correction of Median Link Relatives of Monthly Pig-Iron Production, 1904-1914	296
101. Calculation of Standard Deviation of the Monthly Items of a Time Series	309
102. Irregular Fluctuations of Values of Building Permits Issued for Twenty Leading Cities, July, 1903-June, 1916	310
103. Calculation of Crude Measure of Correspondence between Two Time Series	316
104. Calculation of Correlation Coefficient from Paired Data in "Cycle" Form	319
105. Calculation of Correlation Coefficient from Time Series — First Method	321
106. Calculation of Correlation Coefficient from Time Series — Second (or Ayres) Method	323
107. Coefficients of Correlation between Cycles of Pig-Iron Production and Interest Rate on Sixty-to-Ninety Day Commercial Paper, with 0, 3, 4, 5, 6, 7, 8, 9, and 12 Months' Lag of Interest Rate	325

TABLES

xxi

TABLE	PAGE
108. Average Farm Prices and Index Numbers of Average Farm Prices for Five Cereals in the United States, 1910-1920	330
109. Monthly Indices of the Average Cost of Living in the United States, 1914-1923	331
110. Monthly Index of Employment in Thirteen Groups of Industries in the United States, 1919-1923	332
111. Monthly Index of Production in Basic Industries in the United States, 1914-1924	333
112. Monthly Index of the Volume of Trade in the United States, 1919-1923	333
113. Weights of Members of Foot-Ball Team before and after Important Contest	343
114. Annual Average Prices and Price Relatives of Group of Nine Commodities, 1913 and 1920	346
115. Illustration of Bias of Arithmetic Mean of Price Relatives	346
116. Average Prices and Price Relatives of Three Common Articles of Consumption, Years A and B	349
117. Average Farm Price on December 1, and Total Production of, Five Cereal Crops in the United States, 1914 and 1919	352
118. Average Monthly Consumption and Monthly Average Prices of Foodstuffs Consumed by Typical Family in 1913 and 1920	354
119. Calculation of Index Number of Cost of Living for Fixed Bill of Goods, 1913 and 1920 (Aggregative Form)	356
120. Calculation of Index Number of Cost of Living for Variable Bill of Goods, 1913 and 1920 (Aggregative Form)	357
121. Percentage Distribution of "Value Added by Manufacture" by Major Groups of Industries	366
122. Seasonal Variation of Price of Hides, 1890-1913	372
123. Bad Debt Losses in Nebraska Retail Stores, 1923	375
124. Frequency Distribution of the Percentage of Members Unemployed, July 1, 1920, in 120 Reporting Trade Unions	394

CHARTS

CHART	PAGE
1. Percentage Distribution of Average Number of Wage Earners in Major Industrial Groups of Manufacturing Establishments in the United States in 1919	53
2. Percentage of National Income Contributed by Various Industries in the United States (Average 1909 to 1918)	54
3. Values of Orchard Fruit Crops in the United States in 1919	55
4. Construction of Steam Vessels 100 Gross Tons and over in the United States, United Kingdom, and Other Countries, in the Fiscal Years 1913 and 1918	56
5. Percentage Distribution of Grades in Large Introductory Courses at Harvard University	61
6. Frequency Distribution of Percentage of Members Unemployed, July 1, 1920, among 120 Reporting Trade Unions	76
7. Percentage Distribution of Weekly Wages among Male Weavers in Cotton Mills of States A and B	79
8. Cumulative Percentage Distribution of Weekly Wages among Male Weavers in Cotton Mills of State A	83
9. Cumulative Percentage Distributions of Length of Service among Employee Exits in Firms A and B	85
10. Lorenz Curves Contrasting Distributions of Wealth in Years A and B	86
11. Lorenz Curve of Wage Earners and Wage Receipts in Given Wage-Earning Group	87
12. Lorenz Curves of Rents and Incomes in Selected Cities	88
13. Population of the United States, January 1, 1920, by States	96
14. Number of Horses on Farms, April 15, 1910, by States	98
15. Number of Sheep on Farms, April 15, 1910, by States	99
16. Geographic Distribution of Corn Production in the United States in 1909	101
17. Geographic Distribution of Expenditures of Farmers in the United States for Fertilizer in 1909	102