

# Epidemiologic Research

PRINCIPLES AND  
QUANTITATIVE METHODS

Kleinbaum  
Kupper  
Morgenstern

# Epidemiologic Research

## PRINCIPLES AND QUANTITATIVE METHODS

David G. Kleinbaum

Lawrence L. Kupper

Hal Morgenstern



LIFETIME LEARNING PUBLICATIONS

A division of Wadsworth, Inc.

Belmont, California

London, Singapore, Sydney, Toronto, Mexico City

*Designer:* Rick Chafan  
*Copy Editor:* Carol Beal  
*Illustrator:* LEPI Graphics  
*Composition:* Interactive Composition Corporation

©1982 by Wadsworth, Inc. All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, Lifetime Learning Publications, Belmont, California 94002, a division of Wadsworth, Inc.

Printed in the United States of America

1 2 3 4 5 6 7 8 9 10—86 85 84 83 82

---

**Library of Congress Cataloging in Publication Data**

Kleinbaum, David G.

Epidemiologic research.

Includes bibliographies and index.

1. Epidemiology—Research. 2. Epidemiology—Research—Methodology. I. Kupper, Lawrence L. II. Morgenstern, Hal. III. Title.

[DNLM: 1. Epidemiologic methods. 2. Models, Theoretical. WA 950 K64e]

RA652.4.K56 614.4'072 82-79

ISBN 0-534-97950-5

AACR2

---

# **Epidemiologic Research**

---

# Table List

- Table 1.1** Code Book for Evans City Data 10
- Table 2.1** Disease Chronicity: Examples by Latency and Duration 23
- Table 2.2** Strategies of Prevention 23
- Table 2.3** Hypothetical Cross-Classification of True Disease Status ( $Y$ ) by Category of a Possible Causal Factor ( $X$ ) 27
- Table 3.1** Types and Objectives of Epidemiologic Research 41
- Table 5.1** Commonly Used Designs for Selected Study Objectives 89
- Table 6.1** Estimation of the Risk ( $R$ ) of Developing Disease X for the Hypothetical Cohort of 12 Subjects (Figure 6.1), by 1-Year Interval ( $t_{j-1}, t_j$ ), Using the Actuarial Method (Equations 6.5 and 6.6) 106
- Table 6.2** Estimation of the Risk ( $R$ ) of Developing Disease X for the Hypothetical Cohort of 12 Subjects (Figure 6.1), by 1-Year Interval, Using the Density Method (Expressions 6.7 and 6.8) 109
- Table 6.3** Number of Prevalent Cases ( $C$ ) and Incident Cases ( $I$ ) During a 2-Year Follow-up of a Hypothetical Stable Population (Size  $N$ ), by Exposure Status and Age ( $j = 1, \dots, 4$ ) 110
- Table 6.4** Estimation of Average Incidence Rates (ID) and Risks ( $R$ ), Using the Density Method (Expressions 6.7 and 6.8), for the Data in Table 6.3, by Exposure Status and Age 111
- Table 7.1** Estimation of Average Incidence Rates (ID), Using the Density Method (Expressions 6.7 and 6.8), and Prevalences ( $P$ ) for the Data in Table 6.3, by Exposure Status and Age 119
- Table 7.2** Estimation of the Risk ( $R$ ) of Dying with Disease X for the Hypothetical Cohort of 12 Subjects (Figure 6.1), by 1-Year Interval ( $t_{j-1}, t_j$ ), Using the Actuarial Method (Expressions 6.5 and 6.6) and the Density Method (Expressions 6.7 and 6.8) 125

<b>Table 7.3</b>	Estimation of the Risk ( $R$ ) of Dying from Any Cause for the Hypothetical Cohort of 12 Subjects (Figure 6.1), by 1-Year Interval ( $t_{j-1}, t_j$ ), Using the Actuarial Method (Expressions 6.5 and 6.6) and the Density Method (Expressions 6.7 and 6.8)	126
<b>Table 7.4</b>	Steady-State Relationships Between Selected Frequency Measures and Incidence, by Frequency Measure and Type of Disease	128
<b>Table 8.1</b>	Data Layouts for Studies Involving Population-Time	141
<b>Table 8.2</b>	Data Layouts for Studies Involving Case-Noncase Status	142
<b>Table 8.3</b>	Estimation of Average Incidence Rates (ID) and Risks ( $R$ ), Using the Density Method (Equations 6.7 and 6.8), and Prevalences ( $P$ ) for the Data in Table 6.3, by Exposure Status and Age	144
<b>Table 8.4</b>	Number of Subjects, by Exposure Category and Disease Status (Hypothetical Data)	145
<b>Table 9.1</b>	Number of Subjects, by Exposure Category and Disease Status (Hypothetical Data)	163
<b>Table 9.2</b>	Approximate Number ( $I$ ) of New Cases and Incidence Rate ( $ID$ ) of Lung Cancer and Coronary Heart Disease (CHD) for the U.S. Population (of Size $N'$ ) in 1976, by Smoking Status	
<b>Table 9.3</b>	Estimated Measures of Association for Smokers Versus Non-smokers and Potential Impact for the U.S. Population in 1976 (Table 9.2), by Disease	170
<b>Table 9.4</b>	Estimable Parameters ( $X$ ) in Epidemiologic Studies, by Type of Design ( $O$ = Optional, Depending on the Study Design)	171
<b>Table 11.1</b>	Selection Probabilities in Controversy About Estrogen Use ( $E$ ) and Presence of Endometrial Cancer ( $D$ )	199
<b>Table 11.2</b>	Target Population (Prevalent Cases) for a Case-Control Study of Skin Cancer ( $D$ ) and Hypertension ( $E$ )	201
<b>Table 11.3</b>	Expected (Actual) Population from Choice of Controls	202
<b>Table 11.4</b>	Calculation of Expected Number of Hospitalizations	204
<b>Table 11.5</b>	Expected (Actual) Hospital Population	205
<b>Table 11.6</b>	Population Selection Probabilities	205
<b>Table 11.7</b>	Expected Hospital Population When Hospitalization Probabilities for $E$ , $D$ , and $D^*$ Are Each .05	207
<b>Table 11.8</b>	Selection Bias Caused by Loss to Follow-up	208
<b>Table 11.9</b>	Incidence for Target Population: Time 0 to Time 1	209
<b>Table 11.10</b>	Expected Cross-sectional (Actual) Population at Time 1	210
<b>Table 11.11</b>	Actual Population from Follow-up Loss ( $\alpha = 1 = \beta$ , $\gamma = .01 = \delta$ ) in Target Population of Table 11.8a	211

- Table 11.12** Adjusting the Study Population 214
- Table 11.13** Minimum Selection Probabilities,  $p^* = 1/\sqrt{1 + \epsilon/OR}$ , Required for  $|OR^\circ - OR| \leq \epsilon$  215
- Table 12.1** Misclassification Bias: Target Versus Actual Populations 222
- Table 12.2** Target Population (No Misclassification) for Example 12.1 224
- Table 12.3** Classification Tables for Example 12.1 225
- Table 12.4** Actual Population (from Misclassification of Disease Status) for Example 12.1 225
- Table 12.5** Target Population (No Misclassification) for Example 12.2 226
- Table 12.6** Classification Tables for Example 12.2 227
- Table 12.7** Actual Population (from Misclassification of Exposure Status) for Example 12.2 227
- Table 12.8** Probabilities of Misclassification for Disease and Exposure 229
- Table 12.9** Correcting for Misclassification in Study Data 232
- Table 12.10** Misclassification Probabilities for Target Population and Study Population in a Case-Control Study 233
- Table 12.11** Fourfold Tables for Target and Actual Populations from the Case-Control Study Described by Table 12.10 233
- Table 12.12** Follow-up Study Sample Data for Example 12.3 235
- Table 12.13** Case-Control Study Sample Data for Example 12.4 236
- Table 13.1** Summary of Examples Showing Confounding and/or Interaction in Randomly Sampled Data 246
- Table 13.2** Examples Showing Confounding and/or Interaction with Randomly Sampled Data 248
- Table 13.3** Subgroup of Examples Illustrating Conditions for Confounding by Study Design 251
- Table 14.1** Joint Confounding in the Risk Ratio Without Marginal Confounding (Two Extraneous Variables) 269
- Table 14.2** Conditional Relationships of Risk Factors  $f$  and  $g$  with  $e$  and  $d$  for the Data in Table 14.1 272
- Table 14.3** Confounding due to Only One Variable When There Are Two Risk Factors 274
- Table 14.4** Conditional Relationships of the Risk Factors  $f$  and  $g$  with  $e$  and  $d$ , Based on Table 14.3 276
- Table 15.1** Data Layout for a Density Follow-up Study 285
- Table 15.2** Density Type of Data from an Obesity Study 287

<b>Table 15.3</b>	Data Layout for a Cumulative Follow-up Study	289
<b>Table 15.4</b>	Revised Data Layout for Table 15.3	289
<b>Table 15.5</b>	Data Layout with Fixed Marginal Totals and with $A = j$	290
<b>Table 15.6</b>	Cumulative Type of Data from the Evans County Heart Disease Study	293
<b>Table 15.7</b>	Data Layout for a Case-Control Study	294
<b>Table 15.8</b>	Results from a Hypothetical Case-Control Study Using Randomly Sampled Cases and Controls	295
<b>Table 15.9</b>	Data Layout for Exact Confidence Intervals for the OR	303
<b>Table 15.10</b>	Case-Control Data	305
<b>Table 16.1</b>	Options for Control of Extraneous Variables	317
<b>Table 17.1</b>	Data Layout for the $g$ th Stratum ( $g = 1, 2, \dots, G$ )	322
<b>Table 17.2</b>	Examples Pertaining to Decisions About Interaction in Stratified Analysis	323
<b>Table 17.3</b>	Crude Analysis of CAT-CHD Data	325
<b>Table 17.4</b>	Summary of Stratified Analyses for CAT-CHD Association	326
<b>Table 17.5</b>	Stratifying on CHL, HPT, and AGE Separately	329
<b>Table 17.6</b>	Risk Ratios from Stratifying on CHL, AGE, and HPT Simultaneously	330
<b>Table 17.7</b>	Stratifying on AGE and ECG to Assess CAT-CHD Association in a Follow-up Study	334
<b>Table 17.8</b>	Example of Inappropriate Use of MH Test (Follow-up Study)	335
<b>Table 17.9</b>	Density Data from Obesity Study Stratified by Attained Age	338
<b>Table 17.10</b>	Stratified Data Involving a Zero Cell Frequency	344
<b>Table 17.11</b>	Relative Weights for Adjusted Risk and Odds Ratios for the Data of Table 17.7	348
<b>Table 17.12</b>	Hypothetical Case-Control Data on Smoking (SMK) and Lung Cancer (LC), Stratified by Age	348
<b>Table 17.13</b>	Adjusted Odds Ratios and Their Relative Weights for the Data of Table 17.12	349
<b>Table 17.14</b>	Adjusted Incidence Density Ratios for the Data of Table 17.9	350
<b>Table 17.15</b>	Data Layout for Stratum $g$ in a Stratified Analysis Involving $c$ Exposure Categories	355
<b>Table 17.16</b>	Formulas for Point and Interval Estimation of Overall Effect in Stratified Analysis	359
<b>Table 18.1</b>	Hypothetical Population	383



- Table 18.2** Follow-up Study to Estimate Risk Ratio; Expected Cell Counts 385
- Table 18.3** Statistics Related to the Estimation of Risk Ratio 385
- Table 18.4** Follow-up Study Using Pair Matching; Expected Numbers of Pairs (Rounded to Integers) 388
- Table 18.5** Case-Control Study to Estimate Exposure Odds Ratio; Expected Cell Counts (Rounded to Integers) 390
- Table 18.6** Statistics Related to the Estimation of Exposure Odds Ratio 390
- Table 18.7** Case-Control Study Using Pair Matching; Expected Numbers of Pairs (Rounded to Integers) 392
- Table 18.8** Recommendations for Analysis Based on Validity and Precision, by Study Design, Population Associations, and Method of Sampling 393
- Table 18.9** Summary of Conclusions About Efficiency Based on Comparisons in Table 18.8 (Cost Considerations Ignored) 393
- Table 18.10** Previous History of Induced Abortion in Cases with Tubal Pregnancy and in Matched Controls 395
- Table 19.1** Hypothetical Data 409
- Table 19.2** Relationship Between Disease and Factor *A* at Different Levels of Factor *B* 410
- Table 22.1** Summary of Stratified Analyses for CAT–CHD Association, Controlling for AGE and ECG 460
- Table 22.2** Precision Measured as Reciprocal Confidence Interval Length 461
- Table 22.3** Point Estimation Results; Evans County Data ( $D = \text{CHD}$ ,  $E = \text{CAT}$ , Controlling for AGE and ECG) 462
- Table 22.4** 95% Large-Sample Confidence Intervals; Evans County Data ( $D = \text{CHD}$ ,  $E = \text{CAT}$ , Controlling for AGE and ECG) 463
- Table 22.5** Test Statistics for  $H_0$ : No CAT–CHD Association; Evans County Data ( $D = \text{CHD}$ ,  $E = \text{CAT}$ , Controlling for AGE and ECG) 464
- Table 22.6** Data Set for Log-Linear Model Analyses Using GSK Weighted Least Squares Estimation 469
- Table 22.7** Strata Resulting from Categorization of Confounder Summarization Scores, Using Four Methods for Forming Strata; Evans County Data ( $D = \text{CHD}$ ,  $E = \text{CAT}$ , Controlling for AGE and ECG) 472
- Table 22.8** Summary Information for Confounder Summarization, Using Methods 1 and 2 for Categorization of CS score ( $D = \text{CHD}$ ,  $E = \text{CAT}$ , Controlling for AGE and ECG) 474
- Table 23.1** Assessing Interaction in the Evans County Data Set When

Controlling for AGE, CHL, SMK, ECG, and HPT, Using Unconditional Maximum Likelihood Estimation of the Logistic Model 479

**Table 23.2** Forward Selection ( $\alpha = .10$ ) Using Strategy 4 with Unconditional ML Estimation 481

**Table 23.3** Best Model Based on Strategy 3;  $-2 \ln \hat{L} = 342.32$  483

**Table 23.4** Estimated Adjusted Odds Ratios for CAT-CHD Association at Various Combinations of Values of Modifiers, Based on Models Determined via Strategies 3 and 4 484

**Table 23.5** Estimated Variance-Covariance Matrix of Estimated Coefficients for Strategy 4 Model Containing CHL and HPT as Effect Modifiers 486

**Table 23.6** 95% Large-Sample Confidence Intervals for ROR(adj.) Based on Strategy 4 Model 487

**Table 23.7** Monitoring Coefficients Involving CAT for Various Subsets of Main Effect (i.e.,  $V_i$ ) Variables, Based on Strategy 4 Model 490

**Table 23.8** Comparison of Adjusted Odds Ratios When AGE, ECG, and SMK Are Retained Versus Being Removed from the Strategy 4 Model 491

**Table 24.1** Results from Unconditional and Conditional ML Estimation of the Logistic Model Adjusted Odds Ratio When Analyzing a Matched Case-Control Study Concerning the Relationship Between Previous Induced Abortion and Subsequent Occurrence of a Tubal Pregnancy 495

**Table 24.2** Distribution of Matched Sets by Smoking Status for a Hypothetical 2-to-1 Matched Case-Control Study Concerning Smoking and Myocardial Infarction 497

**Table 24.3** Summary Statistics for SMK, SBP, and ECG for Cases and Controls Based on a Hypothetical 2-to-1 Matched Case-Control Study Concerning Smoking and Myocardial Infarction 497

**Table 24.4** Hypothetical Data Set for a 2-to-1 Matched Case-Control Study Concerning Smoking and Myocardial Infarction 498

**Table 24.5** Testing for Interaction Effects in a Logistic Model Using Conditional ML Estimation in the Second Matching Example 501

**Table 24.6** Point Estimation and Tests of Significance for the Logistic Model Adjusted Odds Ratio, Controlling for Various Subsets Among SBP and ECG in the Second Matching Example 502

**Table B.1** Standard Normal Cumulative Probabilities 513

**Table B.2** Percentiles of the Chi-square Distribution 516

# Preface

This text discusses the principles and methods involved in the planning, analysis, and interpretation of epidemiologic research studies. Our purpose in writing this book is to provide the applied researcher with a synthesis of current methodological thought and practice. In particular, we focus our discussion on quantitative (including statistical) issues that arise during the course of an epidemiologic investigation. Instead of restricting our focus just to statistical techniques and their various applications, we also address issues of study *design*, *measurement*, and *validity*. Where appropriate, we describe statistical considerations relevant to each such issue.

**AIM**

Throughout the text, we emphasize that validity should be the primary goal in an epidemiologic study, even if this means sacrificing generalizability. A study free of bias, even if restricted in scope, is preferred to a more general study with unresolvable validity problems. Sophisticated statistical analyses mean nothing if the data are unreliable. Or, to put it another way, “good statistical analyses do not salvage poor data.”

**PHILOSOPHY  
AND  
APPROACH**

For several of the topics we discuss (e.g., the treatment of confounding, the use of matching, multivariable analyses), there are differing points of view in the literature. Even when there is reasonable consensus, there are rarely clear-cut recommendations that can be made. Thus, rather than try to find the nonexistent “ideal” solution, we provide a quantitative framework through which such issues can be investigated. Whenever possible, we discuss both the advantages and disadvantages of various approaches to a given methodological problem.

## AUDIENCE AND PRE- REQUISITES

This book was written with several audiences in mind. Our primary audience includes epidemiologists and other health professionals, as well as graduate students in the health sciences. We assume that this audience has a good understanding of the basic principles of epidemiology. A second audience includes researchers or students in other human science disciplines. Finally, a third audience consists of statisticians and biostatisticians interested in the application of statistics to epidemiologic research.

The degree of mathematical sophistication required, particularly with regard to statistical prerequisites, will depend on the reader's learning objectives for each chapter. We have written primarily at the level of persons with a good understanding of basic statistical procedures of data analysis typically covered in a two-course sequence in applied biostatistics. It would also be advantageous—through not essential—for the reader to have a basic knowledge of the biostatistical methods used to analyze epidemiologic data. Because we do not draw heavily on statistical inference principles prior to Part III, the reader with minimal statistical knowledge will be able to follow the discussion of basic methodological principles given in Parts I and II.

## OUTLINE OF TEXT

This book is divided into three parts, plus an introductory chapter (Chapter 1). This overview chapter provides a nonquantitative discussion of key methodological issues that are addressed in the remainder of the text.

Part I (Chapters 2–9) contains the epidemiologic foundation for the more quantitatively oriented (including statistical) treatment of methodological issues addressed in Parts II and III. Of the eight chapters in Part I, only Chapters 6, 7, and 8 contain material (including notation) that is absolutely prerequisite for Parts II and III.

Part II (Chapters 10–14) provides a quantitative, conceptual framework for evaluating validity. Part III (Chapters 15–24) provides a detailed quantitative discussion of procedures and strategies for the design and analysis of epidemiologic research studies.

We strongly recommend that the reader work the exercises provided at the end of the chapters. These exercises are intended to help the reader understand and apply the principles and methods presented in the text. Abbreviated answers to selected problems are provided in Appendix A. A detailed solutions manual for the exercises is available from the publisher.

## ACKNOWLEDGMENTS

We wish to acknowledge several people who contributed to the preparation of this text. The authors developed much of their interest in quantitative epidemiology from exposure to the pioneering and sometimes controversial research contributions of Olli Miettinen at Harvard University. They have also gained insight from the work of Norman Breslow and Nicholas Day, as well as from many others too numerous to list individually. We thank Beverly Cole for typing the many drafts and revisions of all chapters, for her assistance in helping us to communicate efficiently

with our publisher, professional colleagues, and students, and for her patience and good cheer throughout this project. Dave and Larry thank Anna Kleinbaum and Cindy Kupper for once again providing much-needed emotional support during the writing of their second book. Hal thanks Cyndie Gareleck for her encouragement and friendship. We thank the many students and colleagues at UNC, Yale, and elsewhere for their helpful comments and interest. Finally, we thank our publisher, Alex Kugushev, for his highly valued enthusiasm, advice, and friendship.

Dave Kleinbaum

Larry Kupper

Hal Morgenstern

*To our parents*

Joslyn and Janet Kleinbaum

Louis and Sylvia Kupper

Joseph and Edith Morgenstern

SPECIAL ACKNOWLEDGMENT

We give special thanks to Sander Greenland of UCLA's School of Public Health for his many contributions: reviewing the manuscript, offering suggestions for improvement, and providing each of us with further insight into many aspects of the material in the text.

# Contents

<b>Table List</b>	<b>xi</b>		
<b>Preface</b>	<b>xvii</b>		
<b>1. Key Issues in Epidemiologic Research: An Overview</b>	<b>1</b>	<b>3. Types of Epidemiologic Research</b>	<b>40</b>
1.0 Preview	2	3.0 Preview	41
1.1 Measuring Disease Frequency	3	3.1 Experiments	42
1.2 Problems of Validity	5	3.2 Quasi Experiments	44
1.3 Problems in Controlling for Extraneous Factors	8	3.3 Observational Studies	47
1.4 Concluding Remarks	13	3.4 Concluding Remarks	48
Notation	13	References	49
References	15		
<b>PART I OBJECTIVES AND METHODS OF EPIDEMIOLOGIC RESEARCH</b>	<b>17</b>	<b>4. Design Options in Observational Studies</b>	<b>51</b>
<b>2. Fundamentals of Epidemiologic Research</b>	<b>19</b>	4.0 Preview	52
2.0 Preview	20	4.1 Subject Selection	52
2.1 Epidemiologic Research	20	4.2 Methods of Observation	55
2.2 Etiologic Research	26	4.3 Concluding Remarks	60
2.3 Concluding Remarks	36	References	61
References	37	<b>5. Typology of Observational Study Designs</b>	<b>62</b>
		5.0 Preview	63
		5.1 Basic Designs	63
		5.2 Hybrid Designs	70
		5.3 Incomplete Designs	77
		5.4 Comparing Observational Designs	88
		5.5 Concluding Remarks	92
		References	93

<b>6. Measures of Disease Frequency: Incidence</b>	<b>96</b>	9.2 Summary of Epidemiologic Measures	169
6.0 Preview	97	9.3 Concluding Remarks	176
6.1 Measures in General	97	Notation	176
6.2 Basic Incidence Measures: Risk and Rate	99	Practice Exercises	177
6.3 Estimation of Average Rates	100	References	180
6.4 Estimation of Risk	103		
6.5 Choosing an Incidence Measure	111	<b>PART II VALIDITY OF EPIDEMIOLOGIC RESEARCH</b>	<b>181</b>
6.6 Concluding Remarks	113		
Notation	114	<b>10. Validity: General Considerations</b>	<b>183</b>
Practice Exercises	114	10.0 Preview	184
References	115	10.1 Brief Literature Review	185
		10.2 Validity and Precision	185
<b>7. Other Measures of Disease Frequency</b>	<b>117</b>	10.3 Definition of Internal Validity	188
7.0 Preview	118	10.4 Direction of Bias	189
7.1 Prevalence Measures	118	10.5 Classification of Bias	190
7.2 Mortality Measures	122	10.6 Concluding Remarks	191
7.3 Age, Period, and Cohort Effects	130	Notation	192
7.4 Concluding Remarks	134	Practice Exercises	192
Notation	135	References	193
Practice Exercises	136		
References	138	<b>11. Selection Bias</b>	<b>194</b>
		11.0 Preview	195
<b>8. Measures of Association</b>	<b>140</b>	11.1 General Formulation	195
8.0 Preview	141	11.2 Direction of Selection Bias	197
8.1 Ratio Measures	143	11.3 Examples of Selection Bias	198
8.2 Difference Measures	150	11.4 Discussion of the Examples	210
8.3 Other Measures of Association	152	11.5 What Can Be Done About Selection Bias?	212
8.4 Concluding Remarks	154	11.6 Concluding Remarks	215
Notation	155	Notation	215
References	157	Practice Exercises	216
		References	218
<b>9. Measures of Potential Impact and Summary of the Measures</b>	<b>159</b>	<b>12. Information Bias</b>	<b>220</b>
9.0 Preview	160	12.0 Preview	221
9.1 Measures of Potential Impact	160		



