Texts in Statistical Science

Statistics for Epidemiology

Nicholas P. Jewell

CHAPMAN & HALL/CRC

Statistics for Epidemiology

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Analysis of Failure and Survival Data
Peter J. Smith

The Analysis and Interpretation of Multivariate Data for Social Scientists
David J. Bartholomew, Fiona Steele, Irini Moustaki, and Jane Galbraith

The Analysis of Time Series — An Introduction, Sixth Edition
Chris Chatfield

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West and J. Harrison

Applied Nonparametric Statistical Methods, Third Edition
P. Sprent and N.C. Smeeton

Applied Statistics — Handbook of GENSTAT Analysis
E.J. Snell and H. Simpson

Applied Statistics — Principles and Examples
D.R. Cox and E.J. Snell

Bayes and Empirical Bayes Methods for Data Analysis, Second Edition
Bradley P. Carlin and Thomas A. Louis

Bayesian Data Analysis, Second Edition
Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin

Beyond ANOVA — Basics of Applied Statistics
R.G. Miller, Jr.

Computer-Aided Multivariate Analysis, Third Edition
A.A. Afifi and V.A. Clark

A Course in Categorical Data Analysis
T. Leonard

A Course in Large Sample Theory
T.S. Ferguson

Data Driven Statistical Methods
P. Sprent

Decision Analysis — A Bayesian Approach
J.Q. Smith

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Elements of Simulation
B.J.T. Morgan

Epidemiology — Study Design and Data Analysis
M. Woodward

Essential Statistics, Fourth Edition
D.A.G. Rees

A First Course in Linear Model Theory
Nalini Ravishanker and Dipak K. Dey

Interpreting Data — A First Course in Statistics
A.J.B. Anderson

An Introduction to Generalized Linear Models, Second Edition
A.J. Dobson

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Introduction to Optimization Methods and their Applications in Statistics
B.S. Everitt

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference
D. Gamerman

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems
V. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Second Edition
D. Collett

Multivariate Analysis of Variance and Repeated Measures — A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics — A Practical Approach
B. Flury and H. Riedwyl

Practical Data Analysis for Designed Experiments
B.S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Statistics for Medical Research
D.G. Altman

Probability — Methods and Measurement
A. O'Hagan

Problem Solving — A Statistician's Guide, Second Edition
C. Chatfield

Randomization, Bootstrap and Monte Carlo Methods in Biology, Second Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Sampling Methodologies with Applications
Poduri S.R.S. Rao

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistical Process Control — Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants, Fourth Edition
S. Letchford

Statistics for Epidemiology
Nicholas P. Jewell

Statistics for Technology — A Course in Applied Statistics, Third Edition
C. Chatfield

Statistics in Engineering — A Practical Approach
A.V. Metcalfe

Statistics in Research and Development, Second Edition
R. Caulcutt

Survival Analysis Using S — Analysis of Time-to-Event Data
Mara Tableman and Jong Sung Kim

The Theory of Linear Models
B. Jørgensen

Statistics for Epidemiology

Nicholas P. Jewell

CHAPMAN & HALL/CRC

A CRC Press Company

Boca Raton London New York Washington, D.C.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20140820

International Standard Book Number-13: 978-1-4822-8601-4 (eBook - PDF)

To Debra and Britta, my very soul of life

Contents

Acknowledgments

The material in this book has grown out of a graduate course in statistical methods for epidemiology that I have taught for more than 20 years in the School of Public Health at Berkeley. I wish to express my appreciation for the extraordinary students that I have met through these classes, with whom I have had the privilege of sharing and learning simultaneously. My thanks also go to Richard Brand, who first suggested my teaching this material, and to Steve Selvin, a lifelong friend and colleague, who has contributed enormously both through countless discussions and as my local S-Plus expert. The material on causal inference depended heavily on many helpful conversations with Mark van der Laan. Several colleagues, especially Alan Hubbard, Madukhar Pai, and Myfanwy Callahan, have selflessly assisted by reading parts or all of the material, diligently pointing out many errors in style or substance. I am forever grateful to Bonnie Hutchings, who prepared the earliest versions of handouts of some of this material long before a book was ever conceived of, and who has been a constant source of support throughout. I also owe a debt of gratitude to Kate Robertus for her incisive advice on writing issues throughout the text.

Finally, my enjoyment of this project was immeasurably enhanced by the love and support of my wife, Debra, and our daughter, Britta. Their presence is hidden in every page of this work, representing the true gift of life.

CHAPTER 1

Introduction

In this book we describe the collection and analysis of data that speak to relationships between the occurrence of diseases and various descriptive characteristics of individuals in a population. Specifically, we want to understand whether and how differences in individuals might explain patterns of disease distribution across a population. For most of the material, I focus on chronic diseases, the etiologic processes of which are only partially understood compared with those of many infectious diseases. Characteristics related to an individual's risk of disease will include (1) basic measures (such as age and sex), (2) specific risk exposures (such as smoking and alcohol consumption), and (3) behavioral descriptors (including educational or socioeconomic status, behavior indicators, and the like). Superficially, we want to shed light on the "black box" that takes "inputs"—risk factors such as exposures, behaviors, genetic descriptors—and turns them into the "output," some aspect of disease occurrence.

1.1 Disease processes

Let us begin by briefly describing a general schematic for a disease process that provides a context for many statistical issues we will cover. Figure 1.1, an adapted version of Figure 2.1 in Kleinbaum et al. (1982), illustrates a very simplistic view of the evolution of a disease in an individual.

Note the three distinct stages of the disease process: *induction, promotion*, and *expression*. The etiologic process essentially begins with the onset of the first cause of the resulting disease; for many chronic diseases, this may occur at birth or during fetal development. The end of the promotion period is often associated with a clinical diagnosis. Since we rarely observe the exact moment when a disease "begins," induction and promotion are often considered as a single phase. This period, from the start of the etiologic process until the appearance of clinical symptoms, is often called the *latency period* of the disease. Using AIDS as an example, we can define the start of the process as exposure to the infectious agent, HIV. Disease begins with the event of an individual's infection; clinical symptoms appear around the onset and diagnosis of AIDS, with the expression of the disease being represented by progression toward the outcome, often death. In this case, the induction period is thought to be extremely short in time and is essentially undetectable; promotion and expression can both take a considerable length of time.

Epidemiological study of this disease process focuses on the following questions:

- Which factors are associated with the induction, promotion, and expression of a disease? These *risk factors* are also known as *explanatory variables, predictors,*

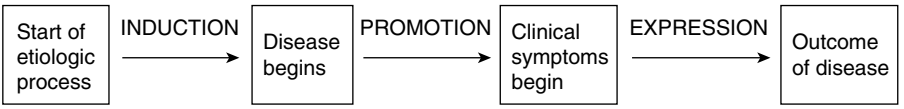

Figure 1.1 *Schematic of disease evolution.*

covariates, independent variables, and *exposure variables*. We will use such terms interchangeably as the context of our discussion changes.

- In addition, are certain factors (not necessarily the same ones) associated with the duration of the induction, promotion, and expression periods?

For example, exposure to the tubercule bacillus is known to be necessary (but not sufficient) for the induction of tuberculosis. Less is known about factors affecting promotion and expression of the disease. However, malnutrition is a risk factor associated with both these stages. As another example, consider coronary heart disease. Here, we can postulate risk factors for each of the three stages; for instance, dietary factors may be associated with induction, high blood pressure with promotion, and age and sex with expression. This example illustrates how simplistic Figure 1.1 is in that the development of coronary heart disease is a continuous process, with no obvious distinct stages. Note that factors may be associated with the outcome of a stage without affecting the duration of the stage. On the other hand, medical treatments often lengthen the duration of the expression of a chronic disease without necessarily altering the eventual outcome.

Disease intervention is, of course, an important mechanism to prevent the onset and development of diseases in populations. Note that intervention strategies may be extremely different depending on whether they are targeted to prevent induction, promotion, or expression. Most public health interventions focus on induction and promotion, whereas clinical treatment is designed to alter the expression or final stage of a disease.

1.2 Statistical approaches to epidemiological data

Rarely is individual information on disease status and possible risk factors available for an entire population. We must be content with only having such data for some fraction of our population of interest, and with using statistical tools both to elucidate the selection of individuals to study in detail (sampling) and to analyze data collected through a particular study. Issues of study design and analysis are crucial because we wish to use sample data to most effectively make applicable statements about the larger population from which a sample is drawn. Second, since accurate data collection is often expensive and time-consuming, we want to ensure that we make the best use of available resources. Analysis of sample data from epidemiological studies presents many statistical challenges since the outcome of interest—disease status—is usually binary. This book is intended to extend familiar statistical approaches for continuous outcome data—for example, population mean comparisons and regression—to the binary outcome context.

1.2.1 Study design

A wide variety of techniques can be used to generate data on the relationship between explanatory factors and a putative outcome variable. I mention briefly only three broad classes of study designs used to investigate these questions, namely, (1) *experimental studies*, (2) *quasi-experimental studies*, and (3) *observational studies*. The crucial feature of an experimental study is the investigator's ability to manipulate the factor of interest while maintaining control of other extraneous factors. Even if the latter is not possible, control of the primary risk factor allows its randomization across individual units of observation, thereby limiting the impact of uncontrolled influences on the outcome. Randomized clinical trials are a type of experimental study in which the main factor of interest, treatment type, is under the control of the investigator and is randomly assigned to patients suffering from a specific disease; other influencing factors, such as disease severity, age, and sex of the patient, are not directly controlled.

Quasi-experimental studies share some features of an experimental study but differ on the key point of randomization. Although groups may appear to differ only in their level of the risk factor of interest, these groups are not formed by random assignment of this factor. For example, comparison of accident fatality rates in states before and after the enactment of seat-belt laws provides a quasi-experimental look at related safety effects. However, the interpretation of the data is compromised to some extent by other changes that may have occurred in similar time periods (did drivers increase their highway speeds once seat belts were required?). A more subtle example involved an Austrian study of the efficacy of the PSA (prostate specific antigen) test in reducing mortality from prostate cancer; investigators determined that, within 5 years, the death rate from prostate cancer declined 42% below expected levels in the Austrian state, Tirol, the only state in the country that offered free PSA screening. Again, comparisons with other areas in the country are compromised by the possibility there are other health-related differences between different states other than the one of interest. Many *ecologic* studies share similar vulnerabilities. The absence of randomization, together with the inability to control the exposure of interest and related factors, make this kind of study less desirable for establishing a causal relationship between a risk factor and an outcome.

Finally, observational studies are fundamentally based on sampling populations with subsequent measurement of the various factors of interest. In these cases, there is not even the advantage of a naturally occurring experiment that changed risk factors in a convenient manner. Later in the book we will focus on several examples including studies of the risk of coronary heart disease where primary risk factors, including smoking, cholesterol levels, blood pressure, and pregnancy history, are neither under the control of the investigator nor usually subject to any form of quasi-experiment. Another example considers the role of coffee consumption on the incidence of pancreatic cancer, again a situation where study participants self-select their exposure categories.

In this book, we focus on the design and analysis of observational epidemiological studies. This is because, at least in human populations, it is simply not ethical to randomly assign risk factors to individuals. Although many of the analytic techniques

are immediately applicable and useful in randomized studies, we spend a considerable amount of effort dealing with additional complications that arise because of the absence of randomization.

1.2.2 Binary outcome data

In studying the relationship between two variables, it is most effective to have refined measures of both the explanatory and the outcome variables. Happily, substantial progress is now being made on more refined assessment of the "quantity" of disease present for many major diseases, allowing a sophisticated statistical examination of the role of an exposure in producing given levels of disease. On the other hand, with many diseases, we are still unable to accurately quantify the amount of disease beyond its presence or absence. That is, we are limited to a simple binary indicator of whether an individual is diseased or not.

Similarly, in mortality studies, while death is a measurable event, the level and quality of health of surviving individuals are notoriously elusive, thus limiting an investigator use of the binary outcome, alive or not. For this reason, we focus on statistical techniques designed for a binary outcome variable. On the other hand, we allow the possibility that risk factors come in all possible forms, varying from binary (e.g., sex), to unordered discrete (e.g., ethnicity), to ordered discrete (e.g., coffee consumption in cups per day), to continuous (e.g., infant birthweight). However, we assume that risk factors or exposures have a fixed value and therefore do not vary over time (although composite values of time-varying measurements, such as cumulative number of cigarette pack-years smoked, are acceptable). Methods to accommodate exposures that change over time, in the context of longitudinal data, provide attractive extensions to the ideas of this book and, in particular, permit a more effective examination of the causal effects of a risk factor. We briefly touch on this again in Chapter 17, and also refer to Jewell and Hubbard (to appear) for an extensive discussion of this topic.

Statistical methodology for binary outcome data is applicable to a wide variety of other kinds of data. Some examples from economics, demography, and other social sciences and public health fields are listed in Table 1.1. In these examples, the nature of a risk factor may also be quite different from traditional disease risk factors.

Table 1.1 *Examples of binary outcomes and associated risk factors*

Binary Outcome	Possible Risk Factors
Use/no use of mental health services in calendar year 2003	Cost of mental health visit, sex
Moved/did not move in calendar year 2003	Family size, family income
Low/normal birthweight of newborn	Health insurance status of mother
Vote Democrat/Republican in 2004 election	Parental past voting pattern
Correct/incorrect diagnosis of patient	Place and type of medical training
Covered/not covered by health insurance	Place of birth, marital status

1.3 Causality

As noted in Section 1.2.1, observational studies preclude, by definition, the randomization of key factors that influence the outcome of interest. This may severely limit our ability to attribute a *causal* pathway between a risk factor and an outcome variable. In fact, selecting from among the three design strategies discussed in Section 1.2.1 hinges on their ability to support a causal interpretation of the relationship of a risk factor or intervention with a disease outcome. This said, most statistical methods are not based, *a priori*, on a causal frame of thinking but are designed for studying associations between factors not necessarily distinguished as input "risk factors" or outcome; for example, the association between eye color and hair color. In short, observational data alone can rarely be used to separate a causal from a noncausal explanation. Nevertheless, we are keenly interested in establishing causal relationships from observational studies; fortunately, even without randomization, there are simple assumptions, together with statistical aspects of the data, that shed light on a putative causal association. In Chapter 8, we introduce much recent work in this regard, including the use of counterfactuals and causal graphs. As noted above, longitudinal observational studies provide greater possibilities for examining causal relationships.

1.4 Overview

Our goal is to introduce current statistical techniques used to collect and analyze binary outcome data (sometimes referred to as *categorical data*) taken from epidemiological studies. The first 11 chapters set the context of these ideas and cover simple methods for preliminary data analysis. Further chapters cover regression models that can be used for the same data. Chapter 16 discusses the special design technique known as *matching* and describes the particular analytic methods appropriate for matched data.

I assume that readers are familiar with basic ideas from a first course in statistics including random variables and their properties (in particular, expectation and variance), sampling, population parameters, and estimation of a population mean and proportion. Of particular importance is the concept of the sampling distribution of a sample estimator that underpins the ideas of interval estimation (confidence intervals) and hypothesis testing (including Type I and Type II errors, p-values, and power). I further anticipate that readers have previously encountered hypothesis tests to compare population means and proportions (for example, the various t-tests and, at least, the one degree of freedom χ^2 test). Familiarity with the binomial, normal, and χ^2 distributions is expected, and experience with the techniques associated with multiple linear regression, while not essential, will make Chapters 12 to 15 much easier to follow. Moore and McCabe (1998) provides an excellent source to review these topics. While mathematical proofs are eschewed throughout, some algebra is used where it can bolster insight and intuition. Fear not, however. Readers are not assumed to have knowledge of techniques that use calculus. The overall goal here is to give some basic driving lessons, not to get under the hood and tinker with the mechanics of the internal combustion engine!

Regression models, found in the second half of the book, can be used to incorporate the simpler analyses from earlier chapters. Some readers may be tempted to jump

directly to these methods, arguing that the earlier stratification methods are really only of historical interest. My view is that basic tabulations with related analyses are important not only to develop a general basis for understanding more complex regression models, but also for gaining a sense of what a particular data set is "saying" before launching a full-scale regression analysis.

I want to say a few words here about developing a personal philosophy about data analysis. Although statistical methodology has an inevitable feel of mathematics, it is more than simply the application of a set of mathematical rules and recipes. In fact, having the perfect recipe is a wonderful advantage, but it does not guarantee a perfect meal. It is crucial that each data analyst construct his own "artistic" principles that can be applied when unraveling the meaning of data. Asking the right questions and having a deep understanding of the context in which a study is designed and implemented are, of course, terrific help as you begin a data analysis. But some general feel for numbers and how to manipulate and illustrate them will also bear considerable fruit. For instance, a sense for the appropriate level in precision in numerical quantities is a valuable tool. As a rough rule of thumb, I do not pay attention to discrepancies between two quantities that are less than 10% of their size. This is not useful in some contexts—knowing a telephone number to within 10% does not get you too far—but in epidemiological studies, this level of difference is often much less than the size of random, let alone other systematic, error. Focusing on such comparisons in the presence of substantial imprecision is putting the priority in the wrong place; it is my statistical version of "don't sweat the small stuff!" Each reader needs a personal style in deciding how best to approach and report a data analysis project. Many other statistical rules of thumb can be found in van Belle (2002), which includes an entire section on epidemiology.

As a brief add-on to the last paragraph, results of data analyses in this book are frequently given as numerical quantities to the second or third decimal place. This is to allow the reader to reconstruct numerical computations and is not meant to reflect how these quantities should be reported in publications or other forms of dissemination.

This book uses a case study approach to illustrate the statistical ideas in ever-expanding generality. Three primary examples are used: the Western Collaborative Group Study of risk factors for coronary heart disease in men (Rosenman et al., 1975), a case-control study of coffee drinking and pancreatic cancer (MacMahon et al., 1981), and a matched pair case-control study of pregnancy and spontaneous abortion in relation to coronary heart disease in women (Winkelstein et al., 1958). There is nothing particular to these choices; most similar examples would be equally effective pedagogically. Because of the use of case studies, the book is intended to be read through chapter by chapter. While much can be gleaned by a quick glance at an isolated chapter, the material is deliberately constructed so that each chapter builds on the previous material.

Analyzing the same examples repeatedly allows us to see the impact of increasingly complex statistical models on our interpretation and understanding of a single data set. The disadvantage is that readers may not be interested in the specific topics covered in these examples and prefer to see the generality of the methods in the context of a wide range of health issues. Therefore, as you follow the ideas, I encourage you to

bring an epidemiological "sketchbook" in which you can apply the ideas to studies of immediate interest, and in which you can note down questions—and perhaps some answers—that arise from your reading of the epidemiological literature. How did an investigator sample a population? How did they measure exposure? How did they deal with other relevant factors? Was matching involved? How were the risk factors coded in a regression model? What assumptions did these choices involve? How is uncertainty reported? What other issues might affect the accuracy of the results? Has causality been addressed effectively?

This is an appropriate time to emphasize that, like most skills, statistical understanding is gained by "doing" rather than "talking." At the end of each chapter, a set of questions is posed to provide readers an opportunity to apply some of the ideas conveyed during the chapter. To return to the analogy above, these assignments give us a chance to take the car out for a spin after each lesson! They should help you differentiate which points you understand from those you might like to review or explore further. They also give illustrative examples that expand on ideas from the chapter.

1.4.1 Caution: what is not covered

The material in this book is loosely based on a one-semester class of interest to beginning graduate students in epidemiology and related fields. As such, the choice of topics is personal and limited. Inevitably, there is much more to implementing, analyzing, and interpreting observational studies than covered here. We spend little or no time on the conceptual underpinnings of epidemiological thinking or on crucial components of many field investigations, including disease registries, public databases, questionnaire design, data collection, and design techniques. Rothman and Greenland (1998) provides a superb introduction to many of these topics.

There are also many additional statistical topics that are not explored in this book. We will not spend time on the appropriate interpretation of p-values, confidence intervals, and power. The Bayesian approach to statistical inference, though particularly appealing with regard to interpretation of parameter uncertainty, is not discussed here. Nor will we delve much into such issues as the impact of measurement error and missing data, standardization, sample size planning, selection bias, repeated observations, survival analysis, spatial or genetic epidemiology, meta-analysis, or longitudinal studies. At the end of each chapter, we include a section on further reading that provides extensions to the basic ideas.

1.5 Comments and further reading

The material in this book has been influenced by three excellent books on the analysis of epidemiological data: Fleiss (1981), Breslow and Day (1980), and Rothman and Greenland (1998). Fleiss (1981) covers the material through Chapter 9, but does not include any discussion of regression models. Breslow and Day (1980) is a beautifully written account of all the methods we discuss, albeit targeted at the analysis of case-control studies. The statistical level is higher and assumes a familiarity with likelihood methods and the theory thereof. Rothman and Greenland (1998) provides an overview

of epidemiological methods extending far beyond the territory visited here, but spends less time on the development and underpinnings of the statistical side of the subject.

Hosmer and Lemeshow (2000) discusses in detail the logistic regression model, but does not include simpler stratification analysis techniques. Collett (2002) is at a similarly high level, and both books approach the material for a general rather than epidemiological application. Schlesselman (1982) has considerably more material on other aspects of epidemiological studies, but a slimmer account of the analysis techniques that are the primary focus here. Kleinbaum et al. (1982) covers all the topics we consider and more, and is encyclopedic in its treatment of some of these ideas. In addition to the above books, more advanced topics can be found in Selvin (1996) and Breslow and Day (1987). Two recent books on similar topics are Woodward (1999) and Newman (2001).

A plethora of statistical packages exist that analyze epidemiological data using methods described in this book. Some of these, including SAS®, SPSS®, GLIM®, and BMDP®, are well known and contain many statistical techniques not covered here; other programs, such as EGRET®, are more tailored to epidemiological data. Collett (2002) contains a brief comparison of most of these packages. A free software package, Epi Info 2000, is currently available online from the Centers for Disease Control. In this book, all data analyses were performed using STATA® or S-Plus®. All data sets used in the text and in chapter questions are available online at http://www.crcpress.com/e_products/downloads/.

I love teaching this material, and whenever I do, I take some quiet time at the beginning of the term to remind myself about what lies beneath the numbers and the formulae. Many of the studies I use as examples, and many of the epidemiological studies I have had the privilege of being a part of, investigate diseases that are truly devastating. Making a contribution, however small, to understanding these human conditions is at the core of every epidemiological investigation. As a statistician, it is easy to be immersed in the numbers churned up by data and the tantalizing implications from their interpetation. But behind every data point there is a human story, there is a family, and there is suffering. To remind myself of this, I try to tap into this human aspect of our endeavor each time I teach. One ritual I have followed in recent years is to either read or watch the Pulitzer prize-winning play, *Wit* by Margaret Edson (or *W;t: a Play*, the punctuation being key to the subject matter). A video/DVD version of the film starring Emma Thompson is widely available. The play gives forceful insight to a cancer patient enrolled in a clinical trial, with the bonus of touching on the exquisite poetry of John Donne.

CHAPTER 2

Measures of Disease Occurrence

A prerequisite in studying the relationship between a risk factor and disease outcome is the ability to produce quantitative measures of both input and output factors. That is, we need to quantify both an individual's exposure to a variety of factors and his level of disease.

Exposure measurement depends substantially on the nature of the exposure and its role in the disease in question. For the purposes of this book, we assume that accurate exposure or risk factor measurements are available for all study individuals. See Section 2.3 for literature on exposure assessment. On the other hand, though fine levels of an outcome variable enhance the understanding of disease and exposure associations, most disease outcomes—and all here—are represented as binary; quantifying a continuous level of disease can involve invasive methods and therefore might be impractical or unethical in human studies. As a consequence, epidemiologists are often reduced to assessing an outcome as disease present or absent. As with risk factors, we assume that accurate binary measurements are available for the disease of interest. Underlying this approach is the simplistic assumption that a disease occurs at a single point of time, so that before this time the disease is not present, and subsequently it is. Disease in exposed and unexposed subgroups of a population is usually measured over an interval of time so that disease occurrences can be observed. This allows for a variety of definitions of the amount of disease in subgroups.

In light of these introductory comments, it is important to note that error in measurement of either exposure or disease or both will compromise the statistical techniques we develop. If such errors are present, this effect must be addressed for us to retain a valid assessment of the disease–exposure relationship. Fortunately, substantial advances have been made in this area and, although beyond the scope of this book, we point out available literature when possible.

2.1 Prevalence and incidence

Disease prevalence and incidence both represent proportions of a population determined to be diseased at certain times. Before we give informal definitions, note that the time scale used in either measurement must be defined carefully before calculation of either quantity. Time could be defined as (1) the age of an individual, (2) the time from exposure to a specific risk factor, (3) calendar time, or (4) time from diagnosis. In some applications, a different kind of time scale might be preferred to chronological time; for example, in infectious disease studies, a useful "time" scale is often defined in terms of the number of discrete contacts with an infectious agent or person.

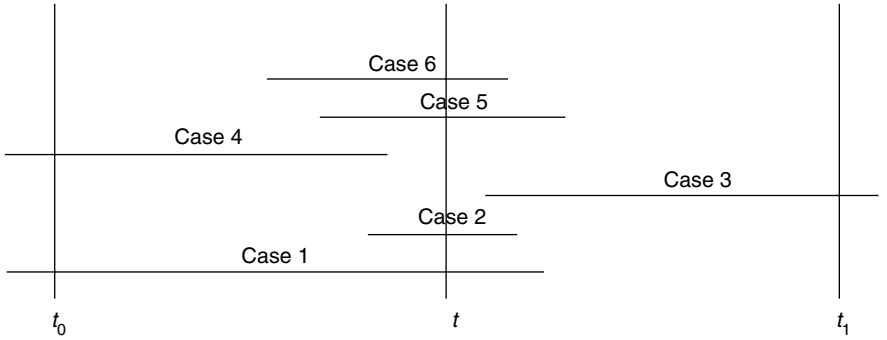

Figure 2.1 *Schematic illustrating calculation of an incidence proportion and point prevalence. Six cases of disease in a population of, say, 100 individuals are represented. Lines represent the duration of disease.*

The *point prevalence* of a disease is the proportion of a defined population at risk for the disease that is affected by it at a specified *point* on the time scale. *Interval prevalence*, or *period prevalence*, is the proportion of the population at risk affected at any point in an interval of time.

The *incidence proportion* is the proportion of a defined population, all of whom are at risk for the disease at the beginning of a specified time interval, who become new cases of the disease before the end of the interval. Since this quantity includes all individuals who become cases over the entire interval, it is sometimes referred to as the *cumulative incidence proportion*. To be "at risk" can mean that an individual has previously been unaffected by the disease, or that susceptibility has been regained after previously contracting the disease and recovering (e.g., as with the common cold to which no sufferer becomes fully immune). There are situations where certain individuals cannot be affected by a disease, e.g., women cannot develop prostate cancer, and so are never at risk.

Figure 2.1 demonstrates schematically the calculation of the incidence proportion and point prevalence in a contrived example of a population of 100 individuals, 6 of whom become cases of a disease during the time period from t_0 to t_1. Using data from the figure, the point prevalence at time t is either $4/100$ or $4/99$, depending on whether case 4 is considered to be at risk of the disease at t or not, respectively. The incidence proportion in the interval $[t_0, t_1]$ is $4/98$, since cases 1 and 4 are not at risk for the disease at the beginning of the interval. This simple scenario reflects that calculations of disease occurrence vary according to definitions of who is "at risk"; take care to compute these quantities according to the appropriate definition!

Neither prevalence (or interval prevalence) nor an incidence proportion carries any units—they are all proportions, sometimes expressed as percentages, that must lie between 0 and 1. The simplest use of these measures of disease occurrence is their comparison across subgroups that have experienced different levels of exposure. For example, one might compare the prevalence of lung cancer among adult males who have smoked at any point in their lives against adult males who have never smoked.

Table 2.1 *Prevalence and incidence data (proportions) on CHD in males*

	Incidence (10 year)		Prevalence	
Cholesterol	CHD	No CHD	CHD	No CHD
High	85 (75%)	462 (47%)	38 (54%)	371 (52%)
Low	28 (25%)	516 (53%)	33 (46%)	347 (48%)

Source: Friedman et al. (1966).

The principal disadvantage with the use of prevalence measures to investigate the etiology of a disease is that they depend not only on initiation, but also on the duration of disease. That is, a population might have a low disease prevalence when (1) the disease rarely occurs or (2) it occurs with higher frequency, but affected individuals stay diseased for only a short period of time (either because of recovery or death). This complicates the role of risk factors, because duration may be influenced by many factors (such as medical treatment) that are unrelated to those that cause the disease in the first place. In addition, risk factors may change during the risk interval and so may assume different values at various times. Thus, prevalence difference across subgroups of a population is often difficult to interpret.

These points are well illustrated by coronary heart disease (CHD), from which a significant proportion of cases has high early mortality. Data (from the Framingham Heart Study, Friedman et al., 1966) relating levels of cholesterol and CHD for men aged 30 to 59 years are shown in Table 2.1. Here, incidence data refer to a group of men, initially free of CHD, whose cholesterol was measured at the beginning of a 10-year follow-up period, during which incident cases of CHD were counted. Cholesterol levels were categorized into four quartiles ("high" and "low" in the table refer to the highest and lowest quartiles). Soon we will discuss methods of analyzing such data with regard to the issue of whether, and by how much, the higher cholesterol group suffers from an elevated risk for CHD. Yet even without analysis, it is immediately clear from the incidence data that there is a substantially larger fraction of CHD cases in the high cholesterol group as compared with the low cholesterol group. This is not apparent in the prevalence data, where cholesterol and CHD measurements were taken at the *end* of the 10-year monitoring period. This discrepancy in the two results might then arise if high cholesterol is associated only with those CHD cases who suffered rapid mortality (dying before the end of the interval) and thus were not included in the prevalence analysis. An alternative explanation is that surviving CHD patients modified their cholesterol levels after becoming incident cases so that their levels at the end of the follow-up period became more similar to the levels of the CHD-free men. (The latter possibility is supported by a more detailed analysis of the Framingham data [Friedman et al., 1966].) This example illustrates the dangers of using prevalence data in attempts to establish a causal association between an exposure and initiation of a disease.

While the statistical methods introduced apply equally to prevalence and incidence data, for most of the discussion and examples we focus on incidence proportions. Why? Because if causality is of prime concern, it is almost always necessary to use incidence, rather than prevalence, as a measure of disease occurrence.

2.2 Disease rates

Before leaving this brief introduction to disease occurrence measures, it is worth broadening the discussion to introduce the concept of a *rate*. If the time interval underlying the definition of an incidence proportion is long, an incidence proportion may be less useful if, for some groups, cases tend to occur much earlier in the interval than for other groups. First, this suggests the need for a careful choice of an appropriate interval when incidence proportions will be calculated. It does not make sense to use an age interval from 0 to 100 years if we want to compare mortality patterns. In the other direction, a study of food-related infections showed much higher mortality effects when individuals were followed for a full year after the time of infection, rather than considering only acute effects (Helms et al., 2003; see Question 5.5). Second, with long risk intervals, there may be substantial variation in risk over the entire period; for example, a person over 65 is at a far higher risk of mortality than an 8 year old. Third, when time periods are long, not all individuals may be at risk over the entire interval; when studying breast cancer incidence, for example, it may make little sense to include premenarcheal childhood years as time at risk. To address the latter point, rates that adjust for the amount of time at risk during the interval are often used.

Specifically, the (average) *incidence rate* of a disease over a specified time interval is given by the number of new cases during the interval divided by the total amount of time at risk for the disease accumulated by the entire population over the same interval. The units of a rate are thus $(\text{time})^{-1}$.

Figure 2.2 is a schematic that allows comparison of the computation of a point prevalence, incidence proportion, and incidence rate. If we assume that the disease under study is chronic in the sense that there is no recovery, then

- The point prevalence at $t = 0$ is $0/5 = 0$; at $t = 5$ it is $1/2 = 0.5$.
- The incidence proportion from $t = 0$ to $t = 5$ is $3/5 = 0.6$.
- The incidence rate from $t = 0$ to $t = 5$ is $3/(5 + 1 + 4 + 3 + 1) = 3/14 = 0.21$ cases per year.

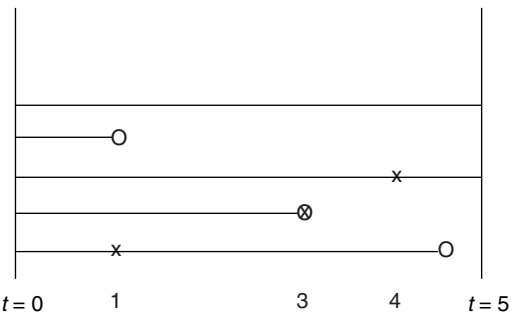

Figure 2.2 *Schematic illustrating calculation of an incidence proportion, point prevalence, and incidence rate. Population of 5; the symbols represent:* O, *death;* x, *incident case of disease. Here, lines represent time alive.*

If population size and follow-up periods are unknown or unclear, or if multiple events per person are possible, the above incidence rate is often referred to by 0.21 cases per person-year, or 0.21 cases per person per year. Note that if the disease is acute, with the property that individuals who recover immediately return to being at risk, then the incidence rate would be $3/(5+1+5+3+4.5) = 0.16$ cases per year.

2.2.1 The hazard function

Returning to the second point of the first paragraph of Section 2.2, if either the population at risk or the incidence rate changes substantially over the relevant time interval, it will be necessary to consider shorter subintervals in order to capture such phenomena. That is, both an incidence proportion—a cumulative measure—and an incidence rate—an average rate over the risk interval—are summary measures for the entire interval and, as such, conceal within-interval dynamics. Unfortunately, limited population size often means that there will be very few incident events in short intervals of time. Nevertheless, in sufficiently large populations it may be possible to measure the incidence rate over smaller and smaller intervals. Such calculations yield a plot of incidence rate against, say, the midpoint of the associated interval on which the incidence rate was based. This kind of graph displays the changes in the incidence rate over time, much as a plot of speed against time might track the progress of an automobile during a journey. This process, in the hypothetical limit of ever smaller intervals, yields the *hazard function,* $h(t)$, which is thus seen as an instantaneous incidence rate.

Figure 2.3 shows a schematic of the hazard function for human mortality among males, where the time variable is age. Looking at mortality hazard curves may feel morbid, particularly for those who find themselves on the right-hand incline of Figure 2.3. However, it is worth remembering that, as Kafka said, the point of life is that it ends. (I have always thought that one of the fundamental points of *The Odyssey* is also that the finiteness of life is what imbues it with meaning.) In Figure 2.3, the hazard function, plotted on the Y-axis, yields the mortality rate (per year) associated with a given age. In a population of size N, a simple interpretation of the hazard function at time t is that the number of cases expected in a *small and unit* increment of time is $Nh(t)$. For example, if $N = 1000$ and the hazard function at time t is 0.005/year, then we roughly anticipate five cases in a year including the time t somewhere near the middle.

If we write the time interval of interest as $[0, T]$, there is a direct link between the hazard function, $h(t)$, for $0 \leq t \leq T$, and the incidence proportion over the interval $[0, t]$, which we denote by $I(t)$. (We assume here that an incident case is no longer at risk after contracting the disease, and that this is the only way in which an individual ceases to be at risk.) The plot of $I(t)$ against t is necessarily increasing—as $I(t)$ measures the cumulative incidence proportion up to time t—and therefore has a positive slope at any time t. It can be shown that

$$h(t) = \frac{dI(t)}{dt} \Big/ (1 - I(t)), \tag{2.1}$$

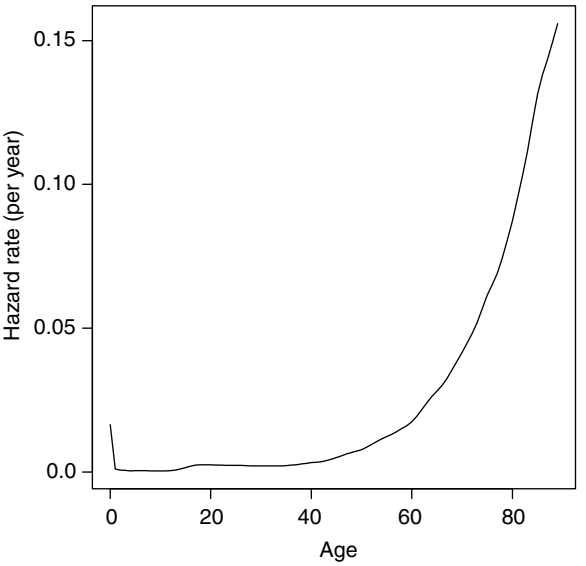

Figure 2.3 *Schematic of the hazard function based on mortality data for Caucasian males in California in 1980.*

where $dI(t)/dt$ represents the slope of $I(t)$ at time t. Note that the term in the denominator accounts for the proportion of the population still at risk at time t. This relationship provides a way of uniquely linking any cumulative incidence proportion, $I(t)$, to a specific hazard function $h(t)$, and vice versa. Oops—I promised no calculus, but this correspondence between incidence and hazard is worth an exception.

When the outcome of interest is disease mortality rather than incidence, we often look at the function $S(t)$, simply defined by $S(t) = 1 - I(t)$. Known as the *survival function*, $S(t)$ measures the proportion of the population that remains alive at age t. Figure 2.4 shows the survival function corresponding to the hazard function of Figure 2.3.

One of the appeals of hazard functions is the ease with which we can extract dynamic information from a plot of the hazard function, as compared with corresponding plots of the incidence proportion or survival function, against time. Even though Figures 2.3 and 2.4 contain exactly the same information, the hazard curve is considerably easier to interpret. For example, to quantify mortality risk for males in the first year of life, observe in Figure 2.3 that this level is roughly the same (yearly) mortality risk faced by 60-year-old males. This comparison is extremely difficult to extract from Figure 2.4. Note the steep increase in mortality risk after age 65 that restricts the chance of extremely long lives. While this phenomenon can be inferred from the graph of the survival function, where a drop occurs at later ages, specific comparative information regarding mortality risks is harder to interpret from the survival function than the hazard function.

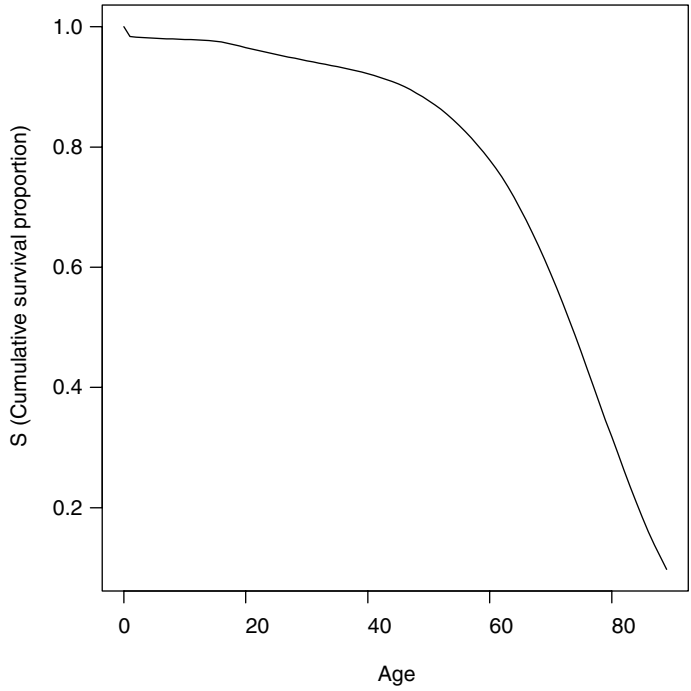

Figure 2.4 *Schematic of the survival function (for mortality) among Caucasian males in California in 1980.*

2.3 Comments and further reading

While we do not consider this matter further here, we cannot understate the value of effective assessment of exposures. With common exposures—tobacco and alcohol consumption, for example—there is substantial literature from previous studies that provides detailed methodology and examples. As a general guide, we refer to the book by Armstrong et al. (1994). In the context of a specific example, there is an excellent discussion of many relevant issues regarding the measurement of exposure to environmental tobacco smoke in the National Academy report on this topic (National Research Council, 1986). For nutritional exposures, the book by Willett (1998) is very useful. Some exposures can only be exactly measured by complex and perhaps invasive procedures so that accurate proxies must often be sought. There is considerable expertise available in questionnaire design and interview procedures to assess exposures determined from a survey instrument.

The definition of an incidence proportion assumes a closed population; that is, no new individuals at risk are allowed to enter after the beginning of the risk period. That this restriction is relaxed when using an incidence rate is one of its principal advantages. That said, it is still possible to estimate an incidence proportion when some individuals enter the population during the risk period. This is sometimes referred

to as delayed entry, or left truncation. An important example arises in studying risks during pregnancy, where the risk period commences at conception but most study participants do not begin observation until a first prenatal visit, or at least until a pregnancy has been detected. The issue of closed vs. open populations will be echoed in our description of study designs in Chapter 5. Variations in incidence rate by exposure are not studied further here, but such data are widely available. The use of Poisson regression models is an attractive approach to rate data, corresponding to the use of regression models for incidence proportions that are studied extensively in Chapters 12 to 15. For further discussion of Poisson regression, see Selvin (1996) and Jewell and Hubbard (to appear).

Comments about exposure assessment are extremely cursory here, and fail to demonstrate many of the complex issues in classifying individuals into differing levels of exposure. Exposure information is often only available in proxy form including self-reports, job records, biomarkers, and ecologic data. To some degree, all exposure measurements are likely to only approximate levels of a true biological agent, even when using standard exposures such as smoking or alcohol consumption histories. Often, it may be valuable to obtain several proxies for exposure, at least on a subgroup of study participants. Validation information—for example, using an expensive but highly accurate exposure measurement on a subset of sampled individuals—is often a crucial component for estimating the properties of exposure measurement error.

In the following chapters, a fixed level of exposure over time is generally assumed. However, chronic exposures, including smoking and occupational conditions, accumulate, implying that the risk of disease will also change over the risk period for this if for no other reason. Composite summary measure of exposure, like pack-years of smoking, should be used with extreme care, since they only capture average exposure information. For example, in studies of fetal alcohol syndrome, episodes of binge drinking may be a better measure of alcohol exposure than average consumption measures. Exposures that vary over time are often best handled with hazard models, which are examined in Chapter 17, where regression methods for incidence proportions are briefly compared with those commonly used for hazard functions. Further details on estimation and inference for hazard functions, and related topics in survival analysis, can be found in Kalbfleisch and Prentice (2002), Hosmer and Lemeshow (1999), and Collett (1994).

2.4 Problems

Question 2.1

Figure 2.5 illustrates observations on 20 individuals in a study of a disease D. The time (i.e., horizontal) axis represents age. The lines, one per individual, represent the evolution of follow-up: the left endpoint signifies the start of follow-up, while the right endpoint indicates the age at onset of D, except in cases of withdrawal for a variety of reasons, marked with a W. For example, the first subject (the lowest line above the axis) started follow-up before his 50th birthday and developed the disease early in his 67th year, i.e., just after turning 66. Calculate for this small population:

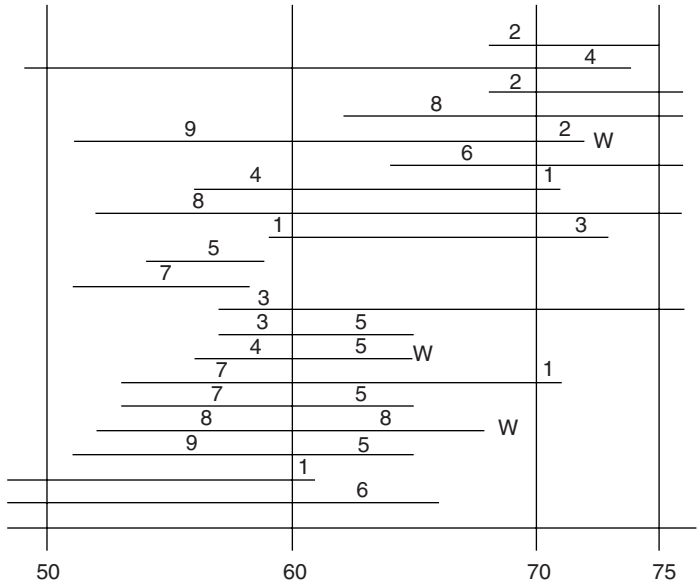

Figure 2.5 *Schematic showing onset of disease at different ages in population of 20 individuals.*

1. The incidence proportion for the disease between the ages 50 and 60 (assume that the disease is chronic so that those with the disease are no longer at risk).
2. The incidence proportion between (a) ages 60 and 70, and (b) ages 70 and 75.
3. The incidence rate for the intervals (a) ages 50 to 60, (b) ages 60 to 70, and (c) ages 70 to 75.

Comment on your findings.

Question 2.2

Indicate whether each of the following computed indices should be considered a point prevalence, incidence proportion, or an incidence rate:

1. The number of children born with congenital heart defects in California in 2002, divided by the number of live births in California in 2002.
2. The number of persons who resided in California on January 1, 2002, and who developed colon cancer during 2002, divided by the total number of disease-free persons who were California residents on January 1, 2002.
3. The number of myopic children under the age of 13 in California on July 1, 2002, divided by the total number of children under the age of 13 in California on July 1, 2002.

4. The number of 60 to 64-year-old California residents who had a stroke in 2002, divided by the total number of 60 to 64-year-old residents on July 1, 2002.

Question 2.3

Describe plausible disease scenarios, with the relevant risk intervals, that suggest (1) an increasing hazard function; (2) a decreasing hazard function; (3) initially increasing hazard, followed by decreasing hazards; and (4) initially decreasing hazard, followed by increasing hazards.

CHAPTER 3

The Role of Probability in Observational Studies

As indicated in the introduction, it is assumed that the reader is familiar with the basic concepts and manipulation of probability and random variables. In this chapter the use of probabilistic terms in epidemiological studies is discussed and some of the simplest ideas are reviewed. One goal of this chapter is to understand what we mean by the risk or probability of a disease. Does it mean that there is some random mechanism inside our bodies that decides our fate? While rejecting that notion, the source of randomness in epidemiological investigations, a key step in quantifying the uncertainty inherent in such studies, is also described. First, some basic understanding of the language and meaning surrounding a probability statement is needed.

Two fundamental components necessary to describe the probability of an occurrence are (1) a *random experiment* and (2) an *event*. A random experiment is a process that produces an identifiable *outcome* not predetermined by the investigator. An event is a collection of one or more distinct possible outcomes. An event occurs if the observed outcome of the experiment is contained in the collection of outcomes defining the event. For example, in tossing a coin one usually thinks of only two possible outcomes—"heads" and "tails." Here, the experiment is the toss of a coin, and an event might be that the coin comes up heads. A qualitatively similar situation occurs with the administration of a specific therapy to a patient with a certain disease. Here, the random experiment is application of treatment, with possible events being "patient cured" or "patient not cured"; note that "not cured" may be defined in terms of combinations of simple outcomes such as blood pressure reading or amount of inflammation.

What is the probability that a tossed coin comes up heads? More generally, in any random experiment, what is the probability of a particular event? Denote the probability of an event A by the term $P(A)$. A heuristic definition of probability is as follows:

> In a random experiment, $P(A)$ is the fraction of times the event A occurs when the experiment is repeated many times, independently and under the exact same conditions.

To express this formally, suppose that a random experiment is conducted K times and the event A occurs in K_A of the total K experiments. As K grows larger and larger, the fraction of times the event A occurs, K_A/K, approaches a constant value. This value is $P(A)$, the probability of A occurring in a single experiment.