Basic Biostatistics for Geneticists and Epidemiologists

A Practical Approach

Robert C. Elston

Department of Epidemiology and Biostatistics, Case Western Reserve University, USA.

William D. Johnson

Pennington Biomedical Research Center, Louisiana State University System, USA.

WILEY

A John Wiley and Sons, Ltd, Publication

Basic Biostatistics for Geneticists and Epidemiologists

Basic Biostatistics for Geneticists and Epidemiologists

A Practical Approach

Robert C. Elston

Department of Epidemiology and Biostatistics, Case Western Reserve University, USA.

William D. Johnson

Pennington Biomedical Research Center, Louisiana State University System, USA.

WILEY

A John Wiley and Sons, Ltd, Publication

This edition first published 2008

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex,
PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Elston, Robert C., 1932–
Basic biostatistics for geneticists and epidemiologists : a practical approach / Robert C. Elston, William D. Johnson.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-02489-8 (hbk) — ISBN 978-0-470-02490-4 (pbk.)
1. Medical statistics. 2. Medical genetics—Statistical methods. 3. Epidemiology—Statistical methods. I. Johnson, William Davis, 1941– II. Title.
[DNLM: 1. Biometry—methods. 2. Data Collection. 3. Epidemiologic Methods. 4. Genetics, Medical. WA 950 E49b 2008]
R853.S7E48 2008
610.72—dc22

2008038609

A catalogue record for this book is available from the British Library

ISBN 978-0-470-02489-8 (Hbk)
ISBN 978-0-470-02490-4 (Pbk)

Set in 11/13.5pt Newcaledonia by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed in the UK by Antony Rowe Ltd, Chippenham, Wiltshire

CONTENTS

PREFACE

'Biostatistics, far from being an unrelated mathematical science, is a discipline essential to modern medicine – a pillar in its edifice' (*Journal of the American Medical Association* (1966) 195: 1145). Today, even more so than forty years ago, anyone who wishes to read the biomedical literature intelligently, especially in the areas of genetics and epidemiology, needs to understand the basic concepts of statistics. It is our hope that this book will provide such an understanding to those who have little or no statistical background and who need to keep abreast of new findings in these two biomedical areas.

Unlike many other elementary books on statistics, the main focus of this book is not so much on teaching how to perform some of the simpler statistical procedures that may be necessary for a research paper, but rather on explaining basic concepts needed to understand the literature. Many of the simpler statistical procedures are in fact described, but computational details are included in the main body of the text only if they help clarify the underlying principles. We have relegated to the Appendix other details that, if included in the body of the text, would tend to make it difficult for the reader to see the forest for the trees. If you wish to have the details, read the notes in the Appendix concurrently, chapter by chapter, with the rest of the book.

This book has been written at an elementary mathematical level and requires no more than high school mathematics to understand. Nevertheless, you may find Chapters 4 and 5 a little difficult at first. These chapters on probability and distributions are basic building blocks for subsequent concepts, however, and you should study them carefully. The basic concepts of estimation and hypothesis testing are covered by the end of Chapter 8, and this is followed in Chapter 9 by some more advanced concepts – but always explained in simple language – that underlie many types of analysis now commonly used by geneticists and epidemiologists. The next three chapters cover special statistical methods that are widely used in both genetic and epidemiological research. There is no attempt, on the other hand, to go into any detail on the advanced statistical methods of analysis used in the special field of genetic epidemiology – this would be a book in itself. In the last chapter we have tried to review the most important concepts introduced in earlier chapters as they relate to a critical reading of reports published in the literature.

We have attempted to illustrate the statistical methods described with enough examples to clarify the principles involved, but without their being so many and so detailed that the reader is caught up in irrelevant and unnecessary technicalities. We have tried to make these examples realistic and yet easy to grasp for someone with a background in human genetics or epidemiology. Because genetic terminology can be confusing to epidemiologists, we briefly introduce the terminology we use in Chapter 1; similarly, for the geneticists, we also give in Chapter 1 a very brief introduction to epidemiology. Apart from providing ideal examples in the application of probability and statistics, genetics is a discipline that underlies all biology, while epidemiology plays a central role in medical research. Detailed knowledge of the molecular aspects of genetics or epidemiology is not, however, necessary to understand the examples.

Each chapter after the first ends with a set of problems and at the end of the book are further review problems. The answers to alternate problems are given at the end of the book.

Robert C. Elston, M.A., Ph.D.
William D. Johnson, Ph.D.

CHAPTER ONE

Key Concepts

deductive reasoning, inductive reasoning
scientific method
statistical inference
variability, reliability of data
population data, population parameter, sample data, sample estimate
autosomes, chromosomes, X chromosome, Y chromosome
genotype, phenotype
alleles, polymorphism, mutation, variant
homozygous, homozygote, heterozygous, heterozygote
locus, loci, diallelic, biallelic, haplotype
epidemic, epidemiology
factors, demographic, economic, genetic, social, temporal
frequency of disease
built environment

Introduction: The Role and Relevance of Statistics, Genetics and Epidemiology in Medicine

WHY BIOSTATISTICS?

In this book on biostatistics we study the application of statistical theory and methods to analyze data collected by geneticists and epidemiologists. Such data are typically collected to further the field of medicine. Accordingly, a genetic study to investigate whether one or more genes might predispose people to an increased risk of developing a specific disease would require an application of statistics to reach a valid conclusion. Similarly, an application of statistics is required to reach a valid conclusion when a clinical study is conducted for the purpose of investigating which of two pharmaceutical treatments is preferred in managing patients with a specific disease. The primary aim of this book is to provide an introduction to statistics with enough detail to address issues such as these but without giving so many mathematical details that the reader loses sight of the end product. We begin by distinguishing between two types of reasoning – inductive reasoning and deductive reasoning. The former is a central theme in the application of statistical inference, but both types of reasoning are used so often in everyday life that it is often difficult to realize that they are really very different from each other.

When taking a clinical history, conducting a physical examination, or requesting laboratory analyses, radiographic evaluations, or other tests, a physician is collecting information (data) to help choose diagnostic and therapeutic actions. The decisions reached are based on knowledge obtained during training, from the literature, from experience, or from some similar source. General principles are applied to the specific situation at hand in order to reach the best decision possible for a particular patient. This type of reasoning – from the general to the specific – is called *deductive*

Basic Biostatistics for Geneticists and Epidemiologists: A Practical Approach R. Elston, W. Johnson

reasoning. Much of basic medical training centers around deductive reasoning. Similarly, much of the training in any basic science is based on general scientific laws and what we can deduce from them.

If it has not happened already, at some point in your training you must ask yourself: How do we obtain the information about what happens in general? A medical student, for example, will learn that patients with hypertension eventually have strokes if their blood pressure is not controlled, but how did we obtain this information in the first place? Does the rule always hold? Are there exceptions? How long can the patient go untreated without having a stroke? Just how high can the blood pressure level be before the patient is in imminent danger? These questions are answered by 'experience'. But how do we pyramid the knowledge we glean from experience so that we do not make the same mistakes over and over again? We save the information gathered from experience and refer to it to make better judgments as we are faced by the need to make new decisions. Moreover, we conduct experiments and comparative studies to focus on questions that arise in our work. We study a few patients (or experimental animals), and from what we observe we try to make rational inferences about what happens in general. This type of reasoning – from the specific subject(s) at hand to the general – is called *inductive reasoning*. This approach to research – pushing back the bounds of knowledge – follows what is known as the *scientific method*, which has four basic steps:

1. Making observations – that is, gathering data.
2. Generating a hypothesis – the underlying law and order suggested by the data.
3. Deciding how to test the hypothesis – what critical data are required?
4. Experimenting (or observing) – this leads to an inference that either rejects or affirms the hypothesis.

If the hypothesis is rejected, then we go back to step 2. If it is affirmed, this does not necessarily mean it is true, only that in light of current knowledge and methods it appears to be so. The hypothesis is constantly refined and tested as more knowledge becomes available.

It would be easy to reach conclusions on the basis of observations, were it not for the variability inherent in virtually all data, especially biological data. Genetics and epidemiology are two basic sciences used in medical research to investigate variability in data in an effort to understand the laws of nature. One of the most common decisions a health professional must make is whether an observation on a patient should be considered normal or abnormal. Is a particular observation more typical of a person with disease or of a person without disease? Is the observation outside the range typically found in a healthy person? If the patient were examined tomorrow, would one obtain essentially the same observation? Obviously, observations such as blood pressure evaluations vary greatly, both at different times on the

same patient and from patient to patient. Clinical decisions must be made with this variability in mind.

Inductive inference is a much riskier procedure than deductive inference. In mathematics, we start with a set of axioms. Assuming that these axioms are true, we use deductive reasoning to prove things with certainty. In the scientific method, we use *inductive inference* and can never prove anything with absolute certainty. In trying to generalize results based on a group of 20 families, you might ask such questions as: If 20 additional families were studied, would the results be very close to those obtained on studying the first 20 families? If a different laboratory analyzed the blood samples, would the results be similar? If the blood samples had been stored at a different temperature, would the results be the same?

WHAT EXACTLY IS (ARE) STATISTICS?

Biostatistics is simply statistics as applied to the biological sciences. A statistic (plural: statistics) is an estimate based on a sample of an unknown numerical quantity in a population, such as the mean height of men age 20. Statistics (singular) is a science that deals with the collection, organization, analysis, interpretation, and presentation of information that can be stated numerically. If the information is based on a sample from a population, we usually want to use this information to make inductive inferences about the population. Perhaps the most difficult aspect of statistics is the logic associated with these inductive inferences, yet all scientific evidence is based on this type of statistical inference. The same logic is used, though not always explicitly, when a physician practices medicine: what is observed for a particular patient is incorporated with what has previously been observed for a large group of patients to make a specific decision about that particular patient. Much of the application of statistical methods centers around using sample data to estimate population parameters such as the population mean, and to test hypotheses about these parameters – such as the hypothesis that two or more populations have identical means. If sample data provide a good representation of the sampled population, then a good application of statistical methods usually leads to good estimates of relevant parameters and good decisions about whether or not certain hypotheses are tenable. As mentioned earlier, however, the obscuring effects of extraneous sources of variability in research data create a difficult environment for making statistical inferences. Statisticians have developed many procedures and formulae for these purposes and they continue to search for methods that provide estimates and statistical tests with improved properties and wider applicability.

Human health appears to be determined largely by genetic predispositions and environmental exposures. New information about medicine and human health is obtained by studying groups of people to investigate their genetic endowment

and environmental conditions that may be in some way linked to their health. Because there is great variability in genetic make-up and environmental exposure in the human population, it is difficult to identify 'silver bullet' treatments for all the many diseases that occur in our population. The problem is further exacerbated by the fact that diseases seldom have a simple etiology, in that there may be multiple causes and promoters of health problems. Despite the obscuring effects of inherent variability and multi-factorial causation, there are many general tendencies that lead to patterns in research data. By investigating these patterns in samples of patients and their families, researchers are able to make inductive inferences about a 'population' of patients to reduce the chance of disease, and to develop and improve disease intervention with the aim of advancing healthy well-being. It is easy to see that sound medical research requires a careful synthesis of expertise in many disciplines, including genetics, epidemiology, and statistics.

REASONS FOR UNDERSTANDING STATISTICS

New scientific knowledge is gained from research. and support for the accuracy of any claim to discovery of new knowledge is almost always gleaned from data that measure investigative outcomes. The scientific pursuit of new wisdom lies in a search for truth. All too often, a line of research takes a turn down a wrong path because a scientist allows his preconceived notions to cloud objectivity. Statistical principles provide an orderly and objective approach to collecting and interpreting research data. In nearly all areas of research, the proper use of statistics is crucial to the validity of the conclusions. Yet many students, especially those in the health professions, tend to avoid learning statistics and some ask: 'Why should I study statistics?' The statement 'If I need a statistician, I will hire one' is also common. But health professionals are frequently faced with data on which they must base clinical judgments. The reliability of the support data from genetic and epidemiological studies plays a fundamental role in making good clinical decisions. You must be able to distinguish between discrepant data and routine variability. As a layperson and as a practitioner, you will be bombarded daily with statistics. To make correct decisions based on the data you have, you must know where those data came from and how they were obtained; you must also know whether conclusions based on those data are statistically valid. Statistics are often misinterpreted, and Disraeli is reputed to have said 'There are lies, damned lies, and statistics' (see Huff, 1954). Hence, there is always a need for the proper use of statistics.

As a scientist you must have an inquiring mind and pursue new ideas with passion, but you must also 'listen'. You must 'listen' to the data and 'hear' what your research outcomes are 'saying'. Most investigators fully understand that if you

use bad experimental technique you may be misguided by faulty outcomes. Many fail to recognize, however, that it is equally important to use good technique and judgment in the statistical analysis in order to reach valid conclusions.

In your scientific development, you will rely heavily on the literature for new information that will change the way you view 'what exactly is knowledge' and the directions that should be taken to further investigate new frontiers. It is important that you be able to read published articles critically. You will need to understand terms such as 'p-value', 'significance level', 'confidence interval', 'standard deviation', and 'correlation coefficient', to mention just a few of the statistical terms that are now common in the scientific literature. This book explains these concepts and puts them to work in strategies that will help you distinguish fact from fancy in everyday life – in newspapers and on television, and in making daily comparisons and evaluations. In addition, it goes beyond a rudimentary introduction and provides the building blocks for developing an understanding of the concepts that may be used in modern genetic and epidemiologic studies. After carefully reading this book, you should have an appreciation of statistics so that you know when, and for what purpose, a statistician should be consulted to raise the level of quality of your research. The vanguard pathway for advancing knowledge rests squarely on the scaffolds of sound research and the ability to communicate the findings of that research effectively so that it is accepted by the scientific community. No matter how eloquent the communiqué, the ultimate merit of new research is judged by (1) the perceived impact factor of the journal it is published in, (2) its subsequent frequency of citation in new peer-reviewed research, and (3) reports of consistent (or inconsistent) findings by other researchers who attempt to replicate original findings when addressing the same issues in their own work. When the findings of a research investigation have a high impact on scientific thinking, they come under the scrutiny of the most outstanding researchers in the area who examine all the strengths and weaknesses, including whether those findings can be independently replicated. Furthermore, you yourself must also be able to understand and evaluate the scientific literature in an intelligent manner. Unfortunately, many of the articles in the medical literature draw invalid conclusions because incorrect statistical arguments are used. Schor and Karten (1966) found that most analytical studies published in well-respected medical journals in 1964 were unacceptable in that the conclusions drawn were not valid in terms of the design of the experiment, the type of analysis performed, or the applicability of the statistical tests used. Unfortunately, things were only slightly better 15 years later. Glantz (1980) reported that about half of the articles published in medical journals that use statistics use them incorrectly. More recently, Ioannidis (2005) investigated articles published in 1990–2003 in three high-impact clinical journals that had been cited in over 1000 other subsequently published peer-reviewed journals. Of 45 highly cited original clinical studies that claimed effective treatment interventions,

results in only 20 were replicated by subsequent investigators, in 7 the same type of effects were found but they were not as strong, in 7 the original results were contradicted, and in 11 replicate studies were never reported. A number of possibilities were posited for the inconsistent or refuted findings, and this opened a debate about the integrity of scientific research and the review process for publication. One area of concern is the lack of excellence in the training of our future generations of researchers. One specific shortcoming was discussed recently by Windish *et al*. (2007). Although physicians rely heavily on findings reported in journal publications and these findings are validated by those reports that support their conclusions through the application of sound statistical principles, the authors concluded that most of the medicine residents studied had insufficient knowledge of statistics to properly interpret the results published in clinical medicine journals.

WHAT EXACTLY IS GENETICS?

Genetics is the study of the transmission of hereditary information from generation to generation. The words '*gene*' and 'genetics' both derive from the same root as the word '*gene*ration'. With rare exceptions, each human cell nucleus contains 46 deeply staining bodies, or *chromosomes,* that carry the hereditary information and, in the strict sense, genetics is the study of how this information is transmitted from parents to offspring. The chromosomes contain the genetic material deoxyribonucleic acid (DNA), and the study of DNA, and how it is transcribed, translated and eventually controls the development of the adult, is often nowadays also considered to be genetics – molecular genetics. As a result, the terminology in genetics is changing fast as more is learned about the processes involved. We therefore summarize here the limited terminology we shall use in this book and how it may differ from what is also commonly seen in the literature.

The concept of the gene is due to Mendel, who used the word 'factor'. He used the word 'factor' in the same way that we might call 'hot' and 'cold' factors, not in the way that we call 'temperature' a factor. In other words, his factor, later called a gene, was the 'level', or specific value, of the genetic factor. In the original terminology, the four blood types A, B, O, and AB are determined by three genes A, B, and O. Nowadays, however, it is common to talk of the ABO gene, and the individual 'levels', A, B, and O, are simply called *alleles,* or *variants,* rather than genes. The genes occur along the chromosomes, which are organized into 22 homologous pairs of *autosomes* and two sex chromosomes, X and Y. Females have two X chromosomes, males have an X and a Y chromosome. Except that the Y chromosome has only a very short segment that is homologous to the X chromosome, the alleles, or genes, similarly occur in pairs at specific locations, or *loci* (singular:

locus) along the chromosomes. Thus, we may talk of a person with AB blood type as having either the A and B genes or the A and B alleles at the ABO locus, which occurs on a particular pair of autosomal chromosomes. To avoid confusion, we shall as much as possible avoid the word 'gene' in this book. A locus will denote the position at which two alleles of a particular gene can occur. If the two alleles are the same, the person is *homozygous* (or a *homozygote*) at that locus; if different, the person is *heterozygous* (or a *heterozygote*). There can be more than two alleles at a particular locus in the population, but only two in each individual: these two alleles comprise the individual's *genotype* at that locus. If, at a particular locus, only two alleles occur in the whole population, the locus is *diallelic* (this is the original term, which we shall use; the etymologically less desirable term 'biallelic' is now often also used).

The A, B, O, and AB blood types are *phenotypes* – what is apparent, the trait that is observed – as opposed to the underlying genotypes, which may or may not be deducible from the phenotype. The B blood type (phenotype), for example, can result from either the BB or the BO genotype. We say that the B allele is *dominant* over the O allele, or equivalently that the O allele is *recessive* to the B allele, with respect to the B blood type (the phenotype). Note that 'dominant' and 'recessive' always denote a relationship between particular alleles with respect to a particular phenotype, though it is not uncommon for one or the other, the alleles or the phenotype, to be implicitly understood rather than explicitly stated. The A, B, O, and AB blood types comprise a *polymorphism*, in the sense that they are alternative phenotypes that commonly occur in the population. Different alleles arise at a locus as a result of *mutation*, or sudden change in the genetic material. Mutation is a relatively rare event, caused for example by an error in replication. Thus the different alleles, alternatives that occur at the same locus, are by origin mutant alleles. Many authors now (incorrectly) use the term 'mutation' for any rare allele, and the term 'polymorphism' for any common allele.

Of the two alleles a person has at each locus, just one is passed on to each offspring. The choice of which allele is transmitted is random, this being Mendel's first law, the law of segregation. Because each offspring has two parents, the number of alleles at each locus is thus maintained at two (except for rare exceptions) in each generation. A *haplotype* is the multilocus analogue of an allele at a single locus, that is, a set of alleles, each from a different locus, that are inherited together from the same parent. A DNA molecule is made up of many thousands of subunits, called *nucleotides*, and a locus originally meant the location of a stretch of hundreds or thousands of such subunits that comprise a gene. A nucleotide that is polymorphic in the population is called a *single nucleotide polymorphism* (SNP, pronounced 'snip'), and the chromosomal location of a SNP is nowadays also often called a locus. To stress that the word 'locus' is being used in its original sense, the location of a sequence of SNPs that form a whole gene, the term *gene-locus* is sometimes used.

WHAT EXACTLY IS EPIDEMIOLOGY?

An epidemic of a disease occurs in a group of people when an unusually large number in the group contract the disease. For many years, the term 'epidemic' (from the Greek, literally, 'upon the population') was used in connection with acute outbreaks such as an unusually large number of people infected with influenza or suffering from vomiting and diarrhea associated with ingestion of contaminated food. However, for some time now it has been used to describe unusual occurrences of chronic health conditions, such as excessive amounts of obesity, heart disease, and cancer in a population. Epidemiology is the study of the frequency of disease occurrence in human populations and subpopulations in search of clues of causation that may lead to prevention and better management of disease. Factors typically used to define subpopulations for epidemiological investigation include: (1) demographic factors such as age, ethnicity, and gender; (2) social factors such as education level, number of people in a household, and religious beliefs; (3) economic factors such as household income, occupation, and value of the home; (4) temporal factors such as birth order, time in years, and season of the year; (5) genetic factors such as might be inferred from parents, sibs, and other relatives; and (6) environmental factors such as related to behavior (e.g. diet, cigarette smoking, and exercise), the built environment (e.g. industrial pollution, densely populated cities and air traffic near large airports), and natural exposures (e.g. radiation from sunlight, pollen from trees and unusual weather).

The occurrence of a disease is related to, or associated with, a factor if the disease is found to occur more frequently in some subpopulations relative to other subpopulations. For example, a condition such as obesity (the response) may be associated with diet (the factor or predictor) if dietary habits and the amount of obesity in some ethnic subpopulations differ from the dietary habits and the amount of obesity in other ethnic subpopulations. Epidemiologists search for causes of disease by studying characteristics of associations between the disease and related factors. The following are characteristic measures of association: (1) Strength of the association – the larger the relative difference in measures of disease among subpopulations that are defined by the levels or categories of a factor, the greater the strength of the association. If we can demonstrate a dose–response type of gradient in the relationship, confidence in the strength of the association is enhanced. (2) Consistency of the association – the association is confirmed in independent studies conducted by different investigators in other populations of subjects. (3) Temporal correctness of the association – exposure to the factor precedes onset of the disease. Alternatively, if we withdraw the exposure in a subpopulation, we may be able to demonstrate a decreased risk of disease relative to a group that continues to be exposed. As researchers confirm these characteristic measures of association in different studies and across different populations, their belief increases that the association may have a causal basis.

HOW CAN A STATISTICIAN HELP GENETICISTS AND EPIDEMIOLOGISTS?

Statistics is a vital component of the research process, from the earliest planning stages of a study to the final presentation of its results. In view of the complexity of many of the statistical methods now used in genetics and epidemiology, and the proliferation of software that has been incompletely tested, the involvement of a statistician in all stages of such a research project is often advisable. This will enhance the efficiency of the study and the scientific credibility of its results. If a study has been improperly planned or executed, no amount of statistical expertise can salvage its results. At the beginning of a study, the statistician's activities might include: (1) recommending study designs to meet the objectives and to increase the amount of information that can be obtained; (2) estimating the number of subjects or families (sample size) required to achieve study objectives; (3) helping develop efficient data-collection forms or websites; and (4) recommending ways to monitor the quality of the data as they are being collected. After the data have been collected and prepared for analysis, the statistician can: (1) recommend the most appropriate methods of analysis, and possibly do the analyses personally if the methods are esoteric; (2) interpret the findings in understandable terms; and (3) review and contribute to the statistical content of any presentations and publications that report the results of the study. Statisticians may also be consulted to help evaluate published papers and manuscripts, or to help prepare sections on experimental design and analysis for grant applications.

Because statistical consultation is a professional collaboration, statisticians should be included in research projects from their beginning. Also, because statistical design and analysis are time-consuming activities, statisticians should be informed well in advance of any deadlines. At the beginning of the study, you should: (1) show the statistician exactly where and how the data will be collected, preferably at the collaboration site; (2) describe the objectives of your study in detail, because they are essential in planning a study design that will extract all pertinent information as efficiently as possible; and (3) inform the statistician of any relevant limitations, such as availability of financing or personnel, which are all factors that must be considered in the study design. To clarify the terminology and basic concepts in a field of interest, the researcher should provide the statistician with background information in the form of basic articles and book chapters in the area being investigated.

Once a strategy for your study has been jointly established, it must be followed carefully, and any necessary changes in procedures must be discussed before they are implemented. Even minor changes may affect the way data are analyzed, and in some instances could invalidate an entire study. Although certain statistical methods

may partially correct mistakes in study design, this is not always possible; it is obviously expedient to avoid making such mistakes at all.

DISEASE PREVENTION VERSUS DISEASE THERAPY

The risk of developing many diseases may be increased significantly by poor behavioral habits and environmental exposures. Obesity and cigarette smoking, for example, have been linked to many health problems. However, genetic factors may predispose to addictive behaviors such as overeating and cigarette smoking. There is growing recognition that the infrastructure of our communities – the man-made or 'built' environment – often influences our safety and health. Thus, researchers continue to elucidate factors that may be linked to increased risk of disease and to suggest possible interventions that may reduce this risk, or successfully manage living with the disease once a person develops it. Medical treatments for disease are continually being developed and improved. It has long been known that many rare diseases are due to variants (mutant alleles) at single genetic loci – the so-called monogenic diseases – and in many cases the environment is also involved. The rare phenotype phenylketonuria, or phenylketones in the urine, for example, is caused by a recessive mutant allele and a diet that includes phenylalanine. Without two mutant alleles, or with a diet deficient in phenylalanine, there is no phenylketonuria. Nowadays studies are being conducted in which the genotypes of hundreds of thousands of SNPs are compared between persons with and without a relatively common disease, such as hypertension or diabetes, in order to determine whether particular genotypes at several loci, or particular genotypes in combination with particular behaviors, predispose to the disease. Any such predisposing genotypes can be used to formulate a predictive genetic test that could be used for personalized medicine.

A FEW EXAMPLES: GENETICS, EPIDEMIOLOGY AND STATISTICAL INFERENCE

Opinion polls. We are all aware of the accuracy of projections made by pollsters in predicting the outcome of national elections before many people have even gone to the polls to cast their vote. This process is based on a relatively small but representative sample of the population of likely voters. It is a classic example of statistical inference, drawing conclusions about the whole population based on a representative sample of that population.

Waiting time. Some years ago the King Tut art exhibit was on display in New Orleans. During the last few days of the exhibition, people waited in line for hours just to enter to see it. On the very last day, the lines were exceptionally long and seemed to be moving terribly slowly. One ingenious man decided to estimate his expected waiting time as follows. He stepped off 360 paces (approximately 360 yards) from his position to the front of the line. He then observed that the line moved 10 paces in 15 minutes. He projected this to estimate a movement of 40 paces per hour or 360 paces in 9 hours. The man then decided that 9 hours was too long a period to wait in line. A man (one of the authors of this book) directly behind this fellow, however, decided to wait and stood in line $9\frac{1}{2}$ hours before seeing the exhibit!

Tuberculosis. About one-third of the world's population is currently infected with tuberculosis, a contagious disease of the lungs that spreads through the air. When infectious people cough, sneeze, talk or spit, they propel tuberculosis bacilli into the air. A person needs only to inhale a small number of bacilli to be infected. Only 5–10% of people who are infected become sick or infectious during their lifetime, suggesting a genetic susceptibility to the disease.

Smoking and disease. Today most health experts believe that smoking is bad for the health – that it increases the risk of diseases such as cancer and heart attacks, and has other deleterious effects on the body. Governments have taken actions to modify or suppress advertising by the tobacco industry, and to educate the public about the harmful effects of smoking. These actions were taken, however, only after many independent studies collected statistical data and drew similar conclusions. Although ignored for many years, it has now been established that nicotine dependence has a genetic component.

Cholesterol and coronary artery disease. High-density lipoprotein (HDL) is a carrier of serum cholesterol. High levels of HDL in the blood seem to be associated with reduced risk of coronary artery disease so that it is called 'good' cholesterol. Lp(a) is a genetic variation of plasma low-density lipoprotein (LDL). A high level of Lp(a) is associated with an increased risk of prematurely developing atherosclerosis. The LDL receptor was discovered by studying the genetics of familial hypercholesteremia.

Diabetes. There are two primary types of diabetes: type I (insulin-dependent or juvenile-onset), which may be caused by an autoimmune response, and type II (non-insulin-dependent or adult-onset). Type I diabetes must be treated with insulin, usually by injection under the skin. Several genetic variants have been discovered that predispose to type II diabetes.

Osteoporosis. Osteoporosis affects both men and women but is more common among women. Although many genetic variants influence bone density in both males and females, at different skeletal sites and in different age groups, it is likely

that the magnitude of individual genetic effects differs in different populations and in different environmental settings.

SUMMARY

1. The scientific method provides an objective way of formulating new ideas, checking these ideas with real data, and pyramiding findings to push back the bounds of knowledge. The steps are as follows: make observations, formulate a hypothesis and a plan to test it, experiment, and then either retain or reject the hypothesis.
2. Sexually reproducing species have somatic cells (body cells), which are diploid [$2n$] (they have two sets of n chromosomes, one from the mother, one from the father) or polyploid [Xn] (they have X sets of n chromosomes), and gametes (reproductive cells) which are haploid [n] (they have only one set of n chromosomes).
3. When parents conceive a child, a single cell is formed. This cell contains a lifetime of information about the offspring provided by two sets of 23 chromosomes for a total of 46 chromosomes. The father contributes one set of 22 autosomes and either a Y or an X chromosome, and the mother contributes a second set of 22 autosomes and an X chromosome.
4. Alleles occur at loci (positions) on a chromosome. If the two alleles at a locus are the same the person is homozygous at that locus, if they are different the person is heterozygous. Alleles can be recessive or dominant with respect to a particular phenotype, in which case the phenotype of the heterozygote is indistinguishable from that of one of the homozygotes.
5. An organism in which both copies of a gene are identical – that is, have the same allele – is said to be homozygous for that gene. An organism that has two different alleles of the gene is said to be heterozygous. Often one allele is *dominant* and the other is *recessive* – the dominant allele will determine which trait is expressed.
6. A mutation is a change in the DNA sequence that occurs by error and a polymorphism is a set of phenotypes with a genetic basis; but these terms are often used to mean a rare allele and a common allele, respectively. A haplotype is the multilocus analogue of an allele at a single locus.
7. An epidemic of a disease occurs in a group of people when an unusually large number in that group, or subpopulation, contract the disease. Factors typically used to define subpopulations for epidemiological investigation include

demographic factors, social factors, economic factors, temporal factors, genetic factors, and environmental factors.

8. Some characteristic indicators that suggest causation are: strength of the association; consistency of the association; and temporal correctness of the association. As researchers confirm these characteristic measures of association in different studies, their belief that the association may be causal increases.

9. Statistics deals with the collection, organization, presentation, analysis, and interpretation of information that can be stated numerically. All data collected from biological systems have variability. The statistician is concerned with summarizing trends in data and drawing conclusions in spite of the uncertainty created by variability in the data.

10. Deductive reasoning is reasoning from the general to the specific. Inductive reasoning is drawing general conclusions based on specific observations. Statistics applies inductive reasoning to sample data to estimate parameters and test hypotheses.

REFERENCES

Glantz, S.A. (1980) Biostatistics: How to detect, correct and prevent errors in the medical literature. *Circulation* 61: 1–7. (You will gain more from this article if you read it after you have learned biostatistics.)

Huff, D. (1954) *How to Lie with Statistics*. New York: Norton. (Everyone who has not read this book should do so: you can read it in bed!)

Ioannidis, J.P.A. (2005) Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* 294: 218–228.

Schor, S.S., and Karten, I. (1966) Statistical evaluation of medical journal manuscripts. *Journal of the American Medical Association* 195: 1123–1128.

Windish, D.M., Huot, S.J., and Green, M.I, (2007) Medicine residents' understanding of the biostatistics and results in the medical literature. *Journal of the American Medical Association* 298: 1010–1022.

CHAPTER TWO

Key Concepts

- cause and effect
- confounding
- target population, study population, study unit, census, parameter
- probability sample:
  - random cluster sample
  - simple random sample
  - stratified random sample
  - systematic random sample
  - two-stage cluster sample
- observational study:
  - cohort/prospective study
  - case–control/retrospective study
  - historical cohort/historical prospective study
  - matched pairs
  - sampling designs
- experimental study:
  - completely randomized
  - fractional factorial arrangement
  - randomized blocks
  - split-plot design
  - changeover/crossover design
  - sequential design
- factorial arrangement of treatments
- response variables, concomitant variables
- longitudinal studies, growth curves, repeated measures studies, follow-up studies
- clinical trial, placebo effect, blinding, masking
- double blinding, double masking
- compliance, adherence
- quasi-experimental studies