Statistics for Biology and Health

Series Editors
M. Gail, K. Krickeberg, J. Samet, A. Tsiatis, W. Wong

Statistics for Biology and Health

Bacchieri/Cioppa: Fundamentals of Clinical Research
Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints
Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS
Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction, 2nd ed.
Gentleman/Carey/Huber/Irizarry/Dudoit: Bioinformatics and Computational Biology Solutions Using R and Bioconductor
Hougaard: Analysis of Multivariate Survival Data
Keyfitz/Caswell: Applied Mathematical Demography, 3rd ed.
Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text, 2nd ed.
Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd ed.
Lange: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.
Manton/Singer/Suzman: Forecasting the Health of Elderly Populations
Martinussen/Scheike: Dynamic Regression Models for Survival Data
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators
Nielsen: Statistical Methods in Molecular Evolution
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software
Proschan/Lan Wittes: Statistical Monitoring of Clinical Trials: A Unified Approach
Siegmund/Yakir: The Statistics of Gene Mapping
Simon/Korn/McShane/Radmacher/Wright/Zhao: Design and Analysis of DNA Microarray Investigations
Sorensen/Gianola: Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics
Stallard/Manton/Cohen: Forecasting Product Liability Claims: Epidemiology and Modeling in the Manville Asbestos Case
Sun: The Statistical Analysis of Interval-censored Failure Time Data
Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model
Ting: Dose Finding in Drug Development
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models
Wu/Ma/Casella: Statistical Genetics of Quantitative Traits: Linkage, Map and QTL
Zhang/Singer: Recursive Partitioning in the Health Sciences
Zuur/Ieno/Smith: Analyzing Ecological Data

David Siegmund
Benjamin Yakir

The Statistics of Gene Mapping

Springer

David Siegmund
Department of Statistics
Stanford University
Stanford, CA 94305
USA
dos@stat.stanford.edu

Benjamin Yakir
Department of Statistics
The Hebrew University of Jerusalem
Jerusalem, Israel 91905
msby@pluto.mscc.huji.ac.il

Library of Congress Control Number: 2006938272

ISBN-10: 0-387-49684-X
ISBN-13: 978-0-387-49684-9
e-ISBN-10: 0-387-49686-6
e-ISBN-13: 978-0-387-49686-3

Printed on acid-free paper.

Printed in the United States of America

9 8 7 6 5 4 3 2 1

springer.com

Dedicated to Lily and Sandra for their love and support

Preface

The ultimate goal of gene mapping is to identify the genes that play important roles in the inheritance of particular traits (phenotypes) and to explain the role of those genes in relation to one another and in relation to the environment. In practice this ambition is often reduced to identifying the genomic neighborhoods where one or a small number of the important genetic contributors to the phenotype are located.

Gene mapping takes place in many different organisms for many different reasons. In humans one is particularly interested in inherited or partly inherited diseases, and hopes that an identification of the responsible genes can lead in the relatively short run to better diagnostics and in the long run to strategies to alleviate the disease. In these cases the phenotype can be qualitative, whether an individual is affected with a particular disease, or it can be quantitative, say the level of a biomarker like cholesterol level, blood pressure, or body mass index, which is known or thought to be related to the disease. In plants or animals gene mapping can be of interest in its own right, or to produce more vigorous hybrid plants of agricultural value or farm animals that are more productive or more disease resistant. It can also involve model organisms, e.g., the plant arabidopsis, inbred strains of mice, or baker's yeast (*S. cerevisiae*), where one hopes to gain basic knowledge yielding insights that are broadly applicable.

The traits, or phenotypes, can be essentially any reproducible quality of an organism, and the goal of the mapping need not even be an actual gene. An example of considerable recent interest is a phenotype measuring the level of expression of some gene or genes under given experimental conditions, with the goal of discovering whether the expression of the gene is controlled by the immediate "upstream" region of the gene (*cis* control) or by some other genomic region, perhaps a master control region for a number of genes that must work in coordination (*trans* control).

There is variability in the expression of essentially any phenotype. There is also variability in the inheritance of genotypes, first because of Mendel's laws, but equally important for gene mapping because of recombination. The level of

variability is such that gene mapping necessarily has a statistical component, which is the subject of this book. At its simplest, gene mapping involves the correlation of the phenotype with the genotype of genetic markers, which are themselves hoped to be located close to, hence correlated with, the genes (or genomic regions) of interest.

Although gene mapping was practiced for a good part of the twentieth century, the subject has changed and grown substantially since the late 1980s. In the mid twentieth century the number of suitable markers was small, on the order of a few per genome (and the genomic location of these markers was often imprecise). As a consequence the principal impediment to gene mapping was the usually large genomic distance between gene and marker, which leads to small correlations between phenotype and marker genotype. The statistical model developed in human genetics to deal with this situation assumed that the mode of inheritance of a trait could be adequately modeled. The model almost invariably involved a single gene with the mode of inheritance, usually dominant or recessive, assumed to be known, and the penetrance, i.e., the conditional probability of expressing the trait given the genotype, also known. The unknown parameter of interest was the genetic distance from gene to marker, as measured by the recombination fraction, which then allowed one to test whether the trait was unlinked (recombination fraction $= 1/2$) or linked (recombination fraction $< 1/2$) and to estimate the recombination fraction.

Since the explosion in the experimental techniques of molecular genetics in the late twentieth century, it has become possible to cover the genome with informative markers. As a consequence genes associated with many simple traits, which generate a large correlation between phenotype and markers that are close to the gene, have been mapped successfully. For complex traits, which may involve multiple genes, it is reasonable to assume that there is some marker close to the gene or genes influencing the trait, but the contribution of any particular gene may be small. This leads to small correlations between marker and phenotype, even if the marker is close to the gene; and this small correlation has now become the primary impediment to successful gene mapping.

The principal goal of this book is to explain the statistical principles of and problems arising in gene mapping. Particular emphasis is placed on the ideas that have arisen with the recent experimental developments leading to the availability of large numbers of molecular markers having known genomic locations and the desire to map progressively more complex traits. Indeed, the so-called "parametric" or "LOD" score method of human genetics, which was the established paradigm in human genetics from the time of the classical paper of N. Morton in 1955 [56] until quite recently and is still frequently used, is mentioned only briefly. (See Ott [57] for a thorough discussion.)

We have attempted to keep the formal statistical and computational requirements for reading the book as few as seems reasonable, with the hope that the book can be understood by a diverse audience. It is not, however, a handbook of methods, but a discussion of concepts, where we assume the

reader wants to play an active role, particularly in performing computational experiments and in thinking about the meaning of the results. Mathematical details are omitted when we did not think they gave added insight into the scientific issues. Since they can range from routine to difficult, we do not recommend that the reader expend effort to fill in the omitted details, unless he or she finds that intellectual activity rewarding for its own sake.

The book is organized globally as follows. The first three chapters deal with basic statistical, computational, and genetic concepts that are used later in the book. The next five are concerned with mapping quantitative traits using data from crosses of inbred strains. Chapters 9–13 involve primarily human genetics. Of those, Chaps. 9 and 11 discuss gene mapping based on data from pedigrees. This is conceptually similar to the first half of the book although necessarily substantially more complicated. Chapters 12 and 13 discuss association analysis, where relations between individuals in pedigrees are replaced by relations in populations. The discussion here is substantially less complete, and is to some extent limited to pointing out certain difficulties arising from complicated and uncertain population history. Chapter 10 involves admixture mapping, which has some features in common with the earlier chapters on gene mapping based on meiotic recombination and others related to population based association analysis.

A more detailed road map is as follows.

Chapter 1 reviews basic statistical concepts that are used throughout the book and explores these concepts computationally. It can be skipped or read quickly by someone familiar with basic statistics and computation in the language `R`.

In Chap. 2, we introduce our basic model relating the phenotype as dependent variable to genotype(s) as independent variable(s). It is a simple linear regression model having its origins in the classical paper of Fisher [30]. A principal attraction of the model is that straightforward variations can be developed to deal with quantitative or qualitative traits, which can be dominant or recessive, and which can involve multiple genes that interact with one another and/or with the environment. These variations are discussed at different places in the book.

Chapter 3 deals with some fundamental concepts of population genetics, and provides an opportunity to introduce some new programming techniques. It contains some difficult material, and except for recombination, which plays a central role throughout the book, most of the chapter is referenced only occasionally. The reader may wish to read Chap. 3 selectively, to get a rough idea of its contents and be prepared to refer to it occasionally.

For the experimental genetics of crosses of inbred lines, one can use the regression model and standard statistical methods to develop direct robust tests for linkage between phenotype and marker genotype. In Chap. 4 we discuss the simplest case of testing a single marker, first when that marker is itself a gene affecting the trait, and then when the marker is linked to a gene affecting the trait; and we see quantitatively the (deleterious) effect of

recombination between gene and marker on the power to determine if the marker is linked.

Since we will usually have no good idea where a gene affecting the trait is likely to be found, in Chap. 5 we introduce the notion of a genome scan, where we test a large number of markers distributed throughout the genome for linkage. This leads naturally to a problem of multiple comparisons that is solved both by computational methods and by theoretical results involving the maximum of a stochastic process. A systematic discussion of the power of a genome scan and of the related idea of confidence intervals for locating a linked gene as precisely as possible is given in Chap. 6.

Chapter 7 introduces in the simple context of experimental genetics the problem of missing information and one simple statistical idea to recapture that information. It turns out that in the context of that chapter, the solution is often more complicated than the problem warrants, but the problem appears again in a more complex form in Chaps. 9, 10, and 13, where missing information poses unavoidable difficulties that require progressively more complex algorithms to address.

Chapter 8 is concerned with more advanced problems in experimental genetics. Our goal here is to introduce the reader to these problems and point out which ones can in principle be dealt with by simple adaptations of methods developed up to that point and which ones pose more serious challenges.

Starting in Chap. 9, we discuss gene mapping in human genetics, where it is intrinsically more complicated, especially when there may be more than one gene and uncontrolled environmental conditions. Our discussion here is less complete than in the first eight chapters. It is essentially limited to pointing out how the theoretical framework of earlier chapters can be adapted to the more complex problems of human genetics and how the problem of missing information, which here moves to center stage, can be addressed. Chapters 9 and 11 are concerned with gene mapping based on inheritance within families. The concepts developed in Chaps. 4–8 are used, somewhat indirectly. Our presentation is designed to bring out the similarities while highlighting important differences. One important conclusion is that because of one's inability to perform breeding experiments in humans, family based methods for gene mapping are intrinsically less powerful than mapping in experimental genetics based on crosses of inbred lines.

Chapters 12 and 13 contain a brief introduction to gene mapping in populations, which is often called association analysis. It has the potential advantage over family based methods of substantially more power, providing one can successfully overcome some potential difficulties arising from the unknown population history. Our discussion is limited to describing a few simple models, the reasons they are attractive, and the potential pitfalls.

Chapter 10 is something of a bridge between the earlier chapters and the last two. While the methods discussed there are very similar to the methods of earlier chapters, and the issue of missing information is closely related to the same issue in Chap. 9, there are also complications of population history.

As indicated above, the classical parametric method of linkage analysis in human pedigrees is discussed only briefly. Also, in choosing to emphasize regression based methods, we have limited our discussion of likelihood methods, which can be very powerful when basic modeling assumptions are satisfied, to cases where they seemed to offer distinct advantages. The modeling assumptions, often in the form that phenotypes are normally distributed, can fail to hold, even approximately. When that happens, likelihood methods can be less robust than regression methods, although no statistical method should be regarded as so automatic that computer output is thought to speak for itself.

Finally, we would like to emphasize that the primary purpose of this book is didactic. The concepts and many related details have been published elsewhere by a large number of authors. We have provided some references to the scientific literature, largely for the purpose of introducing the reader to the substantial primary literature; but we have not tried to provide a complete scholarly bibliography.

For feedback in classes where preliminary versions of this book were used, we would like to thank students at The Hebrew University of Jerusalem, the Weizmann Institute, Stanford University, and the National University of Singapore. We also thank those universities, along with the Free University of Amsterdam, the Israel–U.S. Binational Science Foundation, the NIH, and the U.S. National Science Foundation for their support.

Stanford and Jerusalem, *David O. Siegmund*
October 2006 *Benjamin Yakir*

Contents

List of Notations and Terminology

QTL = Quantitative trait locus.
y = Phenotype.
x = Copy number of a given allele.
α = Additive effect.
δ = Dominance effect.
$\tilde{\alpha}$ = Additive effect in orthogonalized model.
$\tilde{\delta}$ = Dominance effect in orthogonalized model.
σ^2_{A} = Additive variance.
σ^2_{D} = Dominance variance.
σ^2_e = Residual variance.
σ^2_y = Variance of the phenotype.
IBD = Identical by descent.
J = Number of alleles shared IBD.
χ = Indicator of population source.
τ = Position of QTL.
p = Frequency of QTL.
g = Penetrance or the number of generations, depending on the context.
t, s = Positions of genetic markers.
f = Frequency of a genetic marker.
θ = Recombination fraction for a single meiosis.
Δ = Distance between (equally spaced) markers.
$\varphi = 2\theta(1-\theta)$ = Probability that IBD state of two half siblings changes between markers separated by a recombination fraction θ.
ϕ = Probability density function for a standard normal distribution.
Φ = Cumulative distribution function for a standard normal distribution.
μ = Mean value.
ξ = Noncentrality parameter.
I = Indicator of an event.
ν = Function associated with overshoot in approximations.
T = Transition probability matrix.
π = Stationary distribution, written as a column vector.

Part I

Background and Preparations

1

Background in Statistics

Statistics is the science that formalizes the process of making inferences from observations. Basic to this process of formalization is the concept of a statistical model. In general, a statistical model is an attempt to provide a mathematical simplification of the mechanism that produced the observations. Statistical models are useful since they allow investigation and optimization of the process of analyzing the observations in a context that is wider than the context of the outcome of the specific trial that is being analyzed. For example, it opens the door to considerations such as: "Had we had the opportunity to try a specific inferential procedure on other datasets, all generated by the same statistical mechanism as the one we observe, how would our procedure perform on the average? What would be the probability of drawing an incorrect conclusion?" Such questions are impossible to address unless we adopt a wider point of view.

Exploration of the properties of statistical procedures may be conducted using mathematical and/or computational tools. In this book we will use mathematical approximations and computerized Monte-Carlo simulations to explore problems of statistical genetics. Monte-Carlo simulation can help one explore scenarios where variability plays a role. The basic idea behind such simulations is to generate a sequence of random datasets and use them to mimic the actual distribution of repeated sampling of the data in the investigation of the statistical procedure. The simulation in this book will be conducted in the `R` programming environment. We start the chapter with a small introduction to `R` and then proceed with a discussion of statistical models and statistical inference in a general framework.

Once a statistical model is set, quantities that connect the observations and the model can be computed. A central one is the likelihood function. The likelihood function is the probability density function of the observations given the statistical model. It is the key for the derivation of efficient inferential tools. In most cases, the statistical model is not confined to a unique distribution function but can be any member of a given family of such distributions. In such a case it is useful to think of the likelihood function as a function of

the parameters that determine the distribution within the family. Varying the values of the parameters will change the value of the likelihood function according to the probability of the observations under the new distribution.

Throughout this book we will introduce statistical models that may fit different scenarios in statistical genetics. We start this chapter by introducing three basic models, which apply not only in genetics: the normal model, binomial model, and Poisson model. In order to illustrate the basic concepts in statistical inference we will consider statistical testing of hypotheses as our primary example. In this context another distribution, the chi-square distribution, will also be introduced. The properties of a statistical test, e.g., its significance level and power, will be discussed. A brief introduction to regression and the concept of a stochastic process, both given a central role in the book, is provided. Later in the chapter we will examine general approaches for constructing statistical tests.

1.1 Introduction to R

R is a freely distributed software for data analysis. In order to introduce R we quote the first paragraphs from the manual *Introduction to R*, written by W. N. Venables, D. M. Smith, and the R Development Core Team. (The full text, as well as access to the installation of the software itself, are available online at `http://cran.r-project.org/`):

> "R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has
>
> - an effective data handling and storage facility,
> - a suite of operators for calculations on arrays, in particular matrices,
> - a large, coherent, integrated collection of intermediate tools for data analysis,
> - graphical facilities for data analysis and display either directly at the computer or on hardcopy, and
> - a well developed, simple and effective programming language (called S) which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.)
>
> The term environment is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.
> R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis."

The R system may be obtained as a source code or installed using a precompiled code on the Linux, Macintosh, or Windows operating system. Programming in R for this book was carried out under Windows. (A more-detailed explanation regarding the installation of R under Windows may be found at the URL http://www.biostat.jhsph.edu/~kbroman/Rintro/Rwin.html.)

After the installation of R under Windows an icon will be added to the desktop. It is convenient to have a separate working directory for each project. For that, one may copy the R icon into the new directory and set the working directory to be the new directory. (Right-click on the icon and choose Properties. Copy the path of the directory to the "start in:" box of the Shortcut slip. Consequently, the given directory will become the default location where R expects to find input and to where it saves its output.) Double clicking on the icon will set the R system going.

The R language is an interactive expression-oriented programming langauge. The elementary commands may consist of expressions, which are immediately evaluated, printed to the standard output, and lost. Alternatively, expressions can be assigned to objects, which store the evaluation of the expression. In the latter case the result is not printed out to the screen. These objects are accessible for the duration of the session, and are lost at the end of the session, unless they are actively stored. At the end of the session the user is prompted to store the entire workspace image, including all objects that were created during the session. If "Yes" is selected, then the objects used in the current session will be available in the next. If "No" is selected, then only objects from the last saved image will remain.

Commands are separated either by a semi-colon (;) or by a new line. Consider the following example, which you should try typing into the window of the R Console after the ">" prompt:

```
> x <- c(1,2,3,4,5,6)
> x
[1] 1 2 3 4 5 6
```

Note that the first line created an object named x (a vector of length 6, which stores the value 1, ..., 6). In the second line we evaluated the expression x, which printed out the actual values stored in x. In the formation of the object x we have applied the concatenation function "c". This function takes inputs and combines them together to form a vector.

Once created, an object can be manipulated in order to create new objects. Different operations and functions can be applied to the object. The resulting objects, in turn, can be stored with a new name or with the previous name. In the latter case, the content of the object is replaced by the new content. Continue the example:

```
> x*2
[1]  2  4  6  8 10 12
```

```
> x
[1] 1 2 3 4 5 6
> x <- x*2
> x
[1]  2  4  6  8 10 12
```

Observe that the original content of `x` was not changed due to the multiplication by two. The change took place only when we deliberately assigned new values to the object `x` using the assignment operator "`<-`".

Say we want to compute the average of the vector `x`. The function "`mean`" can be applied to produce:

```
> mean(x)
[1] 7
```

A more complex issue is to compute the average of a subset of `x`, say the values larger than 6. Selection of a sub-vector can be conducted via the vector index, which is accessible by the use of square brackets next to the object. Indexing can be implemented in several ways, including the standard indexing of a sequence using integers. An alternative method of indexing, which is natural in many applications, is via a vector with logical `TRUE/FALSE` components. Consider the following example:

```
> x > 6
[1] FALSE FALSE FALSE  TRUE  TRUE  TRUE
> x[x > 6]
[1]  8 10 12
> mean(x[x > 6])
[1] 10
```

The vector created by the expression "`x > 6`", and which is used in the example for indexing, is a logical vector of the same length as the vector `x`. Only the components of `x` having a "`TRUE`" value in the logical indexing vector are selected. In the last line of the example above the resulting object is used as the input to the function "`mean`", which produces the expected value of 10.

For comparison consider a different example:

```
> x*(x > 6)
[1]  0  0  0  8 10 12
> mean(x*(x > 6))
[1] 5
```

In this example we multiplied a vector of integers `x` with a vector of logical values "`x > 6`". The result is a vector of length 6 with zero components where the logical vector takes the value "`FALSE`" and the original values of `x` where the logical value takes the value "`TRUE`". Two points should be noted. Observe that `R` can interpret a product of a vector with integer components and a vector with logical components in a reasonable way. Standard programming

languages may produce error messages in such circumstances. In this case, R translates the logical vector into a vector with integer values – one for "TRUE" and zero for "FALSE". The outcome, a product of two vectors with integer components, is a vector of the same type. A second point to make is that multiplication of two vectors using "*" is conducted term by term. It is not the inner product between vectors. A different operator is used in R in order to compute inner products.

As in any programming langauge, R requires experience – something that can be obtained only though practice. We will base simulations in this book on R. Starting with very simple examples, we will gradually present more sophisticated code. Our hope is that during that process any reader who did not have a previous exposure to R will learn to use the system and will share our appreciation of the beauty of the langauge and its usefulness for conducting simple simulations in statistics. Indeed, for understanding our exposition, an ability to *read* R is more or less necessary. To solve exercises, programs may be written in other programming languages. In the first chapters of the book we do not assume familiarity with R. Thus, detailed explanations will accompany code lines. These explanations will become less detailed as we progress through the chapters. A reader who is interested in a more systematic introduction to the system is encouraged to use any of the many introductory resources to R that can be found in the form of books or online documents. (Consult, for example, the contributed documentation in http://cran.r-project.org/.)

1.2 The Binomial, Poisson, and Normal Models

We now return to the main subject matter of this chapter by considering three popular statistical models: the *binomial*, the *normal*, and the *Poisson* random variables.

The Binomial Model

Assume that the observations can be represented as a sequence of n binary outcomes. In such a case, the possible outcomes may be classified as *success* or *failure*, or numerically coded as 1 or 0. Such a sequence is termed *Bernoulli trials* if the trials are statistically independent of each other (i.e., the probability of success in one trial is not affected by the outcomes in the other trials).

Suppose the probability of success is the same for all trials. Denote this probability by p and let the random variable X denote the total number of successes among the n trials. Then X is said to have a binomial distribution. For future reference, it will be helpful to observe that X can be regarded as the sum of n independent Bernoulli random variables, which are themselves the special case of binomial random variables with $n = 1$. The probability density function of X is given by:

$$f(x) = \Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n \; .$$

The short notation $X \sim B(n, p)$ is used to refer to this distribution. The expectation of X (i.e., the average value of X, denoted "$\mathrm{E}(X)$") is equal to np, and its variance (denoted "$\mathrm{var}(X)$") is equal to $np(1-p)$ (with $[np(1-p)]^{1/2}$ the standard deviation of X).

(A brief summary of some properties of expectations, variances, and covariances can be found at the end of the chapter. While not strictly necessary for what follows, it will facilitate understanding of some calculations that otherwise must be accepted "on faith.")

The Normal Distribution

The normal distribution – also known as the *Gaussian* distribution – is a very popular statistical model. The formula for the density of the normal distribution is given by:

$$f(x) = \frac{\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}, \quad -\infty < x < \infty \; ,$$

which forms the famous bell shape. The parameter μ is the mean, or the location, of the distribution and σ^2 is its variance (σ is the standard deviation or the scale parameter). In particular, when $\mu = 0$ and $\sigma^2 = 1$ the distribution is called the *standard* normal distribution. The density of the standard normal distribution is symbolized by "$\phi(x)$" and the cumulative distribution function (cdf) is symbolized by

$$\Phi(x) = \int_{-\infty}^{x} \phi(z) \mathrm{d}z \; .$$

We denote the fact that X has a normal distribution with mean μ and variance σ^2 by the notation $X \sim N(\mu, \sigma^2)$.

An important property of the normal distribution is that if we add or subtract independent, normally distributed variables, the result is again normally distributed. Symbolically, if X_1 and X_2 are independent with $X_i \sim N(\mu_i, \sigma_i^2)$, then $X_1 \pm X_2 \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$.

Statistics are quantities computed as functions of the observations. The distribution of a statistic can be quite complex. Surprisingly, it is frequently the case that the distribution of a statistic resembles the bell-shaped distribution of the normal random variable, provided that the sample size is large enough and the statistic is computed as an average or a function of averages. One form of this result will be stated more formally later in this chapter when we discuss the *central limit theorem* (CLT).

The Poisson Distribution

The Poisson distribution is useful in the context of counting the occurrences of rare events. Like the binomial distribution, it takes integer values. As we will see later in this chapter, it can arise as an approximation to the binomial distribution when p is small and n is large.

We say that a random variable X has a Poisson distribution with mean value λ (written $X \sim \text{Poisson}(\lambda)$) if the probability function of X has the form

$$f(x) = \mathrm{e}^{-\lambda} \frac{\lambda^x}{x!} , \quad x = 0, 1, 2, \dots .$$

The expectation and the variance of X are both equal to λ.

If X_1 and X_2 are independent and Poisson distributed with parameters λ_1 and λ_2, the sum $X_1 + X_2$ is Poisson distributed with parameter $\lambda_1 + \lambda_2$.

1.3 Testing Hypothesis

Statistical inference is used in order to detect and characterize meaningful signals that may be hidden in a environment contaminated by random noise. Hypothesis testing is a typical step in the process of making inferences. In this step one often tries to answer the simple question: "Is there any signal at all?" In other words, can the observed data be reasonably explained by a model for which there is no signal – only noise?

1.3.1 The Structure of a Statistical Test of Hypotheses

Assuming the statistical model has been set, we describe the process of testing a statistical hypothesis in three steps: (i) formulation of the hypotheses, (ii) specification of the test, and (iii) reaching the final conclusion. The first two steps are carried out on the basis of the statistical model, and in principal can be conducted prior to the collection of the observations. Only the third step involves the actual data.

(i) Formulating the hypotheses: A model corresponds to a family of possible distributions. Some of the distributions describe signals in the data, while others describe only noise (for example, when treatment and control data vary but are on average the same). The set of the distributions where there is no signal is called the *null hypothesis* or "H_0". The collection of distributions containing signals is denoted the *alternative hypothesis* or "H_1".

In many cases the classification into the two possible hypotheses can be based on some parameters. In such a case, we specify the hypotheses by a partition of the space of parameters that determine the distribution into two parts – one representing the null hypothesis and the other representing the alternative. For example, the hypotheses can often be formulated in terms

of the mean value μ of some observations. The null hypothesis may correspond to $\mu = 0$ (denoted "$H_0 : \mu = 0$"). The alternative hypothesis may then correspond to the case where the expectation is not equal to zero. (The alternative in this case is called *two-sided* and it is denoted by "$H_1 : \mu \neq 0$".) In other cases there are scientific reasons why negative expectations cannot occur and the alternative corresponds to positive values of the expectation ("$H_1 : \mu > 0$"). Such an alternative hypothesis is termed *one-sided*.

(ii) Specifying the test: In this step one decides which statistic to use for the test and which values of the statistic should correspond to rejection of the null hypothesis. The selected statistic is called the *test statistic* and the set of values for which the null hypothesis is rejected is called the *rejection region*.

For example, for testing whether the population mean μ is equal to zero the average of the observations can be used as a test statistic. Large values of this statistic indicate that the null hypothesis is not correct when a one-sided alternative is tested. Similarly, large absolute values are an indication for rejection if a two-sided alternative is considered. The error of rejecting the null hypothesis when it is true is called a "Type I" error and is usually considered more serious than the other type of error (failure to reject the null hypothesis when it is false). Consequently, the selection of a threshold for the rejection region is determined from the distribution of the test statistic under the null hypothesis. The probability of a Type I error is called the *significance level* of the test. The threshold is set to meet a required significance level criteria, traditionally taken to be 5%.

(iii) Reaching a conclusion: After the stage is set, all that is left is to carry out the test. The value of the test statistic for the observed data set is computed. This is the *observed value* of the statistic. A verdict is determined based on whether the observed value of the statistic falls inside or outside the rejection region. Failing to reject the null hypothesis often means dropping the line of investigation and looking for new directions. Rejection of the null hypothesis is a trigger for the initiation of more statistical analysis aimed at a characterization of the signal.

1.3.2 Testing Genetic Identity by Descent of Affected Siblings

This book summarizes an array of strategies for learning relations between expressed heritable traits and genes – the carrier of the genetic information for the formation of proteins. One of these strategies, called the *affected sib-pairs* (ASP) approach, calls for the collection of a large number of nuclear families, each with a pair of affected siblings that share the condition under investigation. Chapter 9 goes into details in describing statistical issues related to this design. Here we consider an artificial, but somewhat simpler, scenario where all the sibling pairs are actually half-siblings, who share only one parent in common, and we concentrate on a single gene, which may or

may not contribute to the disease. The aim is to test the null hypothesis of no contribution.

The gene may be embodied in any one of several variant forms, called *alleles*. On autosomal chromosomes an individual carries two homologous copies of the gene, one inherited from the mother and the other from the father. Therefore, each offspring carries two versions of the given gene, which may not be identical in form. Still, one of the genes is an identical copy of one of the two homologous genes in the common parent while the other is a copy of one of the homologous genes in the other parent. Concentrate on the copies in the half-siblings that originated from the common parent. There are two possibilities: both half-siblings' copies emerge from a common ancestral source or else each was inherited from a different source. In the former case we say that the two copies are *identical by descent* (IBD), and in the latter case we say that they are not IBD. It is natural to model the IBD status of a given pair as a Bernoulli trial, with an IBD event standing for success. Counting the number of half-sibling pairs for which a gene is inherited IBD would produce a binomial random variable.

At a locus unrelated to the trait, Mendel's laws governing segregation of genetic material from parent to offspring will produce IBD or not with equal probabilities, since whichever gene the first child inherited, the second child has a 50% chance of inheriting the same gene. This probability of IBD is the probability of success when the gene does not contribute to the trait. Suppose, however, that the gene does contribute to the trait. Since both siblings share the trait one may reasonably expect an elevated level of sharing of genetic material within the pair, thus an elevated probability of IBD. Denote by J the IBD count for a given pair, with $J = 0$ or $J = 1$, and let π be the probability that $J = 1$. A natural formulation of the statistical hypothesis is given by $H_0 : \pi = 1/2$ versus $H_1 : \pi > 1/2$. Given a sample of n pairs of half-siblings who share the trait, one may use as a test statistic the number of pairs that share an allele IBD. Since each pair can be regarded as a Bernoulli trial, if we also assume the parents are unrelated to one another, the trials are independent, so the sum is binomially distributed with paramters n and π. One may standardize this binomially distributed statistic by subtracting out the expectation and dividing by the standard deviation, both computed under the null distribution: $B(n, 1/2)$. The standardized statistic is:

$$Z_n = \frac{\sum_{i=1}^{n} J_i - n/2}{(n/4)^{1/2}} \, .$$

A common recommendation is to reject the null hypothesis if Z_n exceeds a threshold of $z = 1.645$, since according to the central limit theorem (discussed below) this will produce a significance level of about 0.05. Values of the test statistic Z_n above that threshold lead to the rejection of the null hypothesis and to the conclusion that the gene contributes to the trait.

Let us investigate the significance level of the proposed test. Assume that a total of $n = 100$ pairs were collected and that in fact $\pi = 1/2$. Then the results of the test may look like this:

```
> n <- 100
> J <- rbinom(1,n,0.5)
> J
[1] 44
> Z <- (J-n/2)/sqrt(n/4)
> Z
[1] -1.2
> Z > 1.645
[1] FALSE
```

The number of pairs that share an IBD copy of the gene was 44 (which is actually less than the expected value of 50). This result was generated using the function "`rbinom`", which simulates the binomial distribution. The first argument of the function is the number of independent copies to produce; a single copy in our case. The second argument is the number of Bernoulli trials. In this example, the number is `n`, which was assigned a value of 100. The third argument is the probability of success, $\pi = 1/2$. The statistic `Z` was computed by standardizing the statistic `J`, which in this case equals the negative value -1.2. Obviously, the null hypothesis is not rejected. Note that the function "`rbinom`" simulates random occurrences of a binomial random variable. Running the same code again may produce different outcomes.

In order to evaluate the significance level of the test it is not enough to simulate a single trial. Consider the following code:

```
> J <- rbinom(10^6,n,0.5)
> Z <- (J-n/2)/sqrt(n/4)
> mean(Z > 1.645)
[1] 0.044226
```

In this case the function "`rbinom`" produces one million independent copies of the binomial distribution, all stored in a vector `J` of that length. Each of the components of the vector `J` is then standardized in the same way as the single number was in the previous example. The result is the vector `Z`, which contains the standardized values. The last line of code involves an application of the function "`mean`", which computes, as we have previously seen, the average value of its input. Note that the input here is a vector with logical `TRUE/FALSE` components. A component takes the value "`TRUE`" if the null hypothesis is rejected and "`FALSE`" when it is not. When introduced to the function "`mean`", the logical values are translated into numerical values: one for "`TRUE`" and zero for "`FALSE`". As a result, the function "`mean`" produces the relative frequency of rejecting the null hypothesis, which is an approximation of the significance level. Observe, that the resulting number is 0.044226. This is close, but not

identical, to the nominal significance level of 0.05, which was based on the central limit theorem, discussed next.

1.4 Limit Theorems

The rationale behind the selection of 1.645 as a threshold in the test discussed above lies in the similarity between the standardized binomial distribution and the standard normal distribution. The given threshold is the appropriate threshold in the normal case. This similarity is justified by the central limit theorem (CLT). In this section we will formulate (without proof) the CLT in the context of sums of independent and identically distributed (i.i.d.) random variables. Actually, the scope of central limit theorems is much wider. It includes multivariate distributions as well as sums of non-identical and weakly dependent random variables. When rare events are considered, the Poisson distribution may provide a better approximation than the normal. A Poisson limit theorem will be presented here in the context of binomial random variables. Again, generalizations of the basic theorem in various directions exist.

The central limit theorem states that the distribution function of a standardized sum of independent and identically distributed random variables converges to the standard normal distribution function. More precisely (recall that Φ denotes the distribution function of the standard normal distribution):

Central Limit Theorem: Let $X_1, X_2, \ldots$, be a sequence of independent and identically distributed random variables. Denote the expectation of these random variables by μ and the variance by σ^2. Let $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. Consider, for each n, the random variable:

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{(n\sigma^2)^{1/2}} = \frac{(\bar{X} - \mu)}{(\sigma/n^{1/2})} = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} .$$

Then, for any $-\infty < x < \infty$,

$$\lim_{n \to \infty} \Pr(Z_n \leq x) = \Phi(x) .$$

As an example of the application of the central limit theorem consider the binomial distribution. Recall that if $X \sim B(n, p)$, then X can be represented as a sum of n Bernoulli variables. Moreover, it is easy to see that the expectation of each of these Bernoulli variables is p and the variance is $p(1-p)$. Hence the distribution of $Z_n = (X - np)/[np(1-p)]^{1/2}$ can be approximated by the standard normal distribution. In particular, when n is large,

$$\Pr(Z_n > 1.645) \approx 1 - \Phi(1.645) = 0.05 .$$

For the example of the preceding section, where $n = 100$ and $\pi = 1/2$, we can actually calculate the exact probability by using the binomial distribution.

The inequality $Z_n > 1.645$ is equivalent to $J > 50 + 1.645 \times (100/4)^{1/2} = 58.225$. The `R` command "`1 - pbinom(58.225,100,.5)`" shows that the exact probability of this event is 0.044313, which is very close to the simulated value found earlier. In more complicated cases it may not be possible to evaluate a probability of interest exactly, and then simulation and/or approximations like those based on the CLT are especially useful.

The normal approximation to the binomial distribution works best when p is not too close to 0 or to 1. When this is not the case the Poisson approximation will tend to produce better results. We can state the theorem that establishes the Poisson approximation as follows:

> **Poisson Approximation:** Let $X_n \sim B(n, p_n)$ be a sequence of binomial random variables. Assume that the sequence of p_n of success probabilities obeys the relation $np_n \to \lambda$, as $n \to \infty$, where $0 < \lambda < \infty$. Then, for any $x = 0, 1, 2, \ldots$,
>
> $$\lim_{n\to\infty} \Pr(X_n = x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} \ .$$

Note that the requirement $np_n \to \lambda$ is equivalent to stating that the probability of success, p_n, converges to zero at a rate that is inversely proportional to the number n of Bernoulli trials. In practice, we have only a single value of n and of p. The Poisson approximation is appropriate when n is large, p is small, and the product $\lambda = np$ is neither very large nor very small, say about 0.5 or 1 to about 5 or 6.

Let us demonstrate both the normal and the Poisson approximation in the binomial setting. The following lines of code will produce the plot in Fig. 1.1:

```
> n <- 100; p <- 0.5
> X <- rbinom(10^6,n,p)
> Z <- (X-n*p)/sqrt(n*p*(1-p))
> z <- seq(-4,4,by=0.01)
> plot(z,pnorm(z),type="l")
> lines(ecdf(Z))
> x <- z*sqrt(n*p*(1-p)) + n*p
> lines(z,ppois(x,n*p),type="s")
```

The first three lines of code require no explanation. They are essentially identical, with `J` replaced by `X`, to the code that was used in order to generate the distribution of the test statistic.

In the forth line we generate a sequence of numbers, ranging between -4 and 4, in jumps of size 0.01. This sequence is generated with the aid of the function "`seq`". The first argument to the function is the starting point of the sequence and the second argument is the ending point. The third argument is the jump size, and it is introduced using the name of the argument "`by`". The rule in the introduction of arguments to functions is that arguments may be set either by placing them in the same order in which they appear in

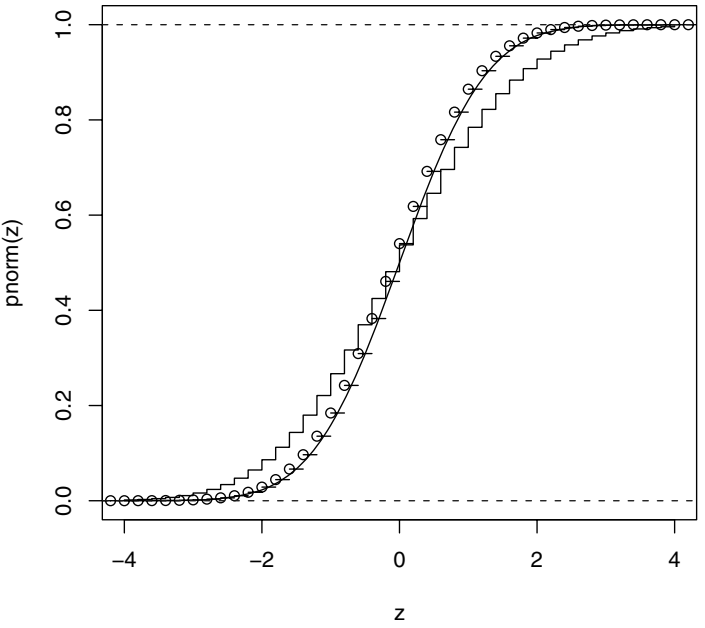

Fig. 1.1. Normal and Poisson approximations: $p = 0.5$, $n = 100$

the definition of the function, or by using the argument assignment format "`par_name = par_value`". If not all preceding arguments are assigned, then the argument must be assigned using the argument assignment format.

The subsequent line produces a plot. The function "`plot`" is a generic function for making plots. In its simplest application it requires as input a sequence of x values and a sequence of y values, both of the same length. It produces the appropriate scatter plot of the points. This basic behavior may be modified by setting arguments. For example, the argument "`type`" determines the plotting style. Setting its value to `"l"` will result in sequentially connecting the points by segments, which will produce a curve. The y values are produced here by the function "`pnorm`". This function takes as input real values and produces as output the normal cumulative distribution function at these values. Execution of the code will result in opening a graphical window within `R` and the generation of the plot of the normal cumulative distribution function over the range of `z`.

The function "`plot`" is classified as a high-level plotting function, since it independently produces a plot. Low-level plotting functions add features to existing plots. The function "`lines`" is a low-level function. In its generic