Chapter 1
Introduction

Chapter Contents

Chapter 1 Introduction

Overview of SAS/Genetics Software

Statistical analyses of genetic data are now central to medicine, agriculture, evolutionary biology, and forensic science. The inherent variation in genetic data, together with the substantial increase in the scale of genetic data following the human genome project, has created a need for reliable computer software to perform these analyses. The procedures offered by SAS/Genetics and described here represent an initial response of SAS Institute to this need.

Although many of the statistical techniques used in the new procedures are standard, others have had to be developed to reflect the genetic nature of the data. All the procedures are designed to operate on data sets that have a familiar structure to geneticists, and that mirror those used in existing software. The syntax for these genetic analyses follows that familiar to SAS users, and the output can be tabular or graphical. The objective of the procedures is to bring the full power of SAS analyses to bear on the characterization of fundamental genetic parameters, and most importantly on the detection of associations between genetic markers and disease status.

Most of the analyses in SAS/Genetics are concerned with detecting patterns of covariation in genetic marker data. These data generally consist of pairs of discrete categories; this pairing derives from the underlying biology, namely the fact that complex organisms have pairs of chromosomes. Each marker refers to the genetic status of a *locus,* each marker type is called an *allele,* and each pair of alleles in an individual is called a *genotype.* A set of alleles present on a single chromosome is called a *haplotype.* Genetic markers may be single nucleotide polymorphisms (SNPs), which are sites in the DNA where the nucleotide varies among individuals, usually with only two alleles possible; microsatellites, which are simple sequence repeats that generate usually between 2 and 20 categories; and other classes of DNA variation.

Two of the procedures in SAS/Genetics are concerned solely with the analysis of genetic marker data. The ALLELE procedure calculates descriptive statistics such as the frequency and variance of alleles and genotypes, as well as estimating measures of marker informativeness, and testing whether genotype frequencies are consistent with Hardy-Weinberg equilibrium (HWE). This procedure also supports three methods for calculation of the degree and significance of *linkage disequilibrium* (LD) among markers at pairs of loci, where LD refers to the propensity of alleles to cosegregate. The HAPLOTYPE procedure is used to infer the most likely multilocus haplotype frequencies in a set of genotypes. Since genetic markers are usually measured independently of one another, there is no direct way to determine which two alleles were on the same chromosome. The algorithm implemented in this procedure converges on the haplotype frequencies that have the highest probability of generating the observed genotypes.

Many genetic data sets are now used to study the relationship between genetic markers and complex phenotypes, particularly disease susceptibility. In general terms, traits can be measured as continuous variables (for example, weight or serum glucose concentration), as discrete numerical categories (for example, meristic measures or psychological class), or as affected/unaffected indicator variables. The two procedures CASECONTROL and FAMILY both take simple dichotomous indicators of disease status and use standard genetic algorithms to compute statistics of association between these indicators and the genetic markers. The CASECONTROL procedure is designed to contrast allele and genotype frequencies between affected and unaffected populations, using three types of chi-square tests and options for controlling correlation of allele frequencies among members of the same subpopulation. Significant associations may indicate that the marker is linked to a locus that contributes to disease susceptibility, though population structure in conjunction with environmental or cultural variables can also lead to associations, and the statistical results must be interpreted with caution. The FAMILY procedure employs several transmission/disequilibrium tests of nonrandom association between disease status and linkage to markers transmitted from heterozygous parents to affected offspring (TDT) or pairs of affected and unaffected siblings (S-TDT and SDT). A joint analysis known as the reconstruction-combined TDT (RC-TDT) can also accommodate missing parental genotypes and families lacking unaffected children under some circumstances.

The output of these procedures can be further explored by using the PSMOOTH procedure to adjust p-values from association tests performed on large numbers of markers obtained in a genome scan, or by creating a graphical representation of the procedures' output, namely p-values from tests for LD, HWE, and marker-disease associations, using the %TPLOT macro.

About This Book

Since SAS/Genetics software is a part of the SAS System, this book assumes that you are familiar with base SAS software and with the books *SAS Language Reference: Dictionary*, *SAS Language Reference: Concepts*, and the *SAS Procedures Guide*. It also assumes that you are familiar with basic SAS System concepts such as creating SAS data sets with the DATA step and manipulating SAS data sets with the procedures in base SAS software (for example, the PRINT and SORT procedures).

Chapter Organization

This book is organized as follows.

Chapter 1, this chapter, provides an overview of SAS/Genetics software and summarizes related information, products, and services. The next five chapters describe the SAS procedures that make up SAS/Genetics software. These chapters appear in alphabetical order by procedure name. They are followed by a chapter documenting a SAS macro provided with SAS/Genetics software.

The chapters documenting the SAS/Genetics procedures are organized as follows:

- The *Overview* section provides a brief description of the analysis provided by the procedure.
- The *Getting Started* section provides a quick introduction to the procedure through a simple example.
- The *Syntax* section describes the SAS statements and options that control the procedure.
- The *Details* section discusses methodology and miscellaneous details.
- The *Examples* section contains examples using the procedure.
- The *References* section contains references for the methodology and examples for the procedure.

Typographical Conventions

This book uses several type styles for presenting information. The following list explains the meaning of the typographical conventions used in this book:

roman — is the standard type style used for most text.

UPPERCASE ROMAN — is used for SAS statements, options, and other SAS language elements when they appear in the text. However, you can enter these elements in your own SAS programs in lowercase, uppercase, or a mixture of the two.

UPPERCASE BOLD — is used in the “Syntax” sections’ initial lists of SAS statements and options.

oblique — is used for user-supplied values for options in the syntax definitions. In the text, these values are written in *italic*.

helvetica — is used for the names of variables and data sets when they appear in the text.

bold — is used to refer to matrices and vectors.

italic — is used for terms that are defined in the text, for emphasis, and for references to publications.

`monospace` — is used for example code. In most cases, this book uses lowercase type for SAS code.

Options Used in Examples

Output of Examples

For each example, the procedure output is numbered consecutively starting with 1, and each output is given a title. Each page of output produced by a procedure is enclosed in a box. Most of the output shown in this book is produced with the following SAS System options:

```
options linesize=80 pagesize=200 nonumber nodate;
```

In some cases, if you run the examples, you will get slightly different output depending on the SAS system options you use and the precision used for floating-point calculations by your computer. This does not indicate a problem with the software. In all situations, any differences should be very small.

Graphics Options

The examples that contain graphical output are created with a specific set of options and symbol statements. The code you see in the examples creates the color graphics that appear in the online (CD) version of this book. A slightly different set of options and statements is used to create the black-and-white graphics that appear in the printed version of the book.

If you run the examples, you may get slightly different results. This may occur because not all graphic options for color devices translate directly to black-and-white output formats. For complete information on SAS/GRAPH software and graphics options, refer to *SAS/GRAPH Software: Reference*.

The following GOPTIONS statement is used to create the online (color) version of the graphic output.

```
filename GSASFILE  '<file-specification>';

goptions gsfname=GSASFILE   gsfmode =replace
         fileonly
         transparency       dev     = gif
         ftext   = swiss    lfactor = 1
         htext   = 4.0pct   htitle  = 4.5pct
         hsize   = 5.625in  vsize   = 3.5in
         noborder           cback   = white
         horigin = 0in      vorigin = 0in ;
```

The following GOPTIONS statement is used to create the black-and-white version of the graphic output, which appears in the printed version of the manual.

```
filename GSASFILE  '<file-specification>';

goptions gsfname=GSASFILE   gsfmode =replace
         gaccess = sasgaedt fileonly
         dev     = pslepsf
         ftext   = swiss    lfactor = 1
         htext   = 3.0pct   htitle  = 3.5pct
         hsize   = 5.625in  vsize   = 3.5in
         border             cback   = white
         horigin = 0in      vorigin = 0in ;
```

In most of the online examples, the plot symbols are specified as follows:

```
symbol1 value=dot color=white height=3.5pct;
```

The SYMBOL*n* statements used in online examples order the symbol colors as follows: white, yellow, cyan, green, orange, blue, and black.

In the examples appearing in the printed manual, symbol statements specify COLOR=BLACK and order the plot symbols as follows: dot, square, triangle, circle, plus, x, diamond, and star.

Where to Turn for More Information

This section describes other sources of information about SAS/Genetics software.

Online Help System

You can access online help information about SAS/Genetics software in two ways. You can select **SAS System Help** from the **Help** pull-down menu and then select **SAS/Genetics Software** from the list of available topics. Or, you can bring up a command line and issue the command **help Genetics** to bring up an index to the statistical procedures, or issue the command **help ALLELE** (or another procedure name) to bring up the help for that particular procedure. Note that the online help includes syntax and some essential overview and detail material.

SAS Institute Technical Support Services

As with all SAS Institute products, the SAS Institute Technical Support staff is available to respond to problems and answer technical questions regarding the use of SAS/Genetics software.

Related SAS Software

Many features not found in SAS/Genetics software are available in other parts of the SAS System. If you do not find something you need in SAS/Genetics software, try looking for the feature in the following SAS software products.

Base SAS Software

The features provided by SAS/Genetics software are in addition to the features provided by Base SAS software. Many data management and reporting capabilities you will need are part of Base SAS software. Refer to *SAS Language Reference: Concepts*, *SAS Language Reference: Dictionary*, and the *SAS Procedures Guide* for documentation of Base SAS software.

SAS DATA Step

The DATA step is your primary tool for reading and processing data in the SAS System. The DATA step provides a powerful general-purpose programming language that enables you to perform all kinds of data processing tasks. The DATA step is documented in *SAS Language Reference: Concepts*.

Base SAS Procedures

Base SAS software includes many useful SAS procedures. Base SAS procedures are documented in the *SAS Procedures Guide*. The following is a list of Base SAS procedures you may find useful:

CHART	for printing charts and histograms
CONTENTS	for displaying the contents of SAS data sets
CORR	for computing correlations
FREQ	for computing frequency crosstabulations
MEANS	for computing descriptive statistics and summarizing or collapsing data over cross sections
PRINT	for printing SAS data sets
SORT	for sorting SAS data sets
TABULATE	for printing descriptive statistics in tabular format
TRANSPOSE	for transposing SAS data sets
UNIVARIATE	for computing descriptive statistics

SAS/GRAPH Software

SAS/GRAPH software includes procedures that create two- and three-dimensional high-resolution color graphics plots and charts. You can generate output that graphs the relationship of data values to one another, enhance existing graphs, or simply create graphics output that is not tied to data.

SAS/IML Software

SAS/IML software gives you access to a powerful and flexible programming language (Interactive Matrix Language) in a dynamic, interactive environment. The fundamental object of the language is a data matrix. You can use SAS/IML software interactively (at the statement level) to see results immediately, or you can store statements in a module and execute them later. The programming is dynamic because necessary activities such as memory allocation and dimensioning of matrices are done automatically. SAS/IML software is of interest to users of SAS/Genetics software because it enables you to program your own methods in the SAS System.

SAS/INSIGHT Software

SAS/INSIGHT software is a highly interactive tool for data analysis. You can explore data through a variety of interactive graphs including bar charts, scatter plots, box plots, and three-dimensional rotating plots. You can examine distributions and perform parametric and nonparametric regression, analyze general linear models and generalized linear models, examine correlation matrices, and perform principal component analyses. Any changes you make to your data show immediately in all graphs and analyses. You can also configure SAS/INSIGHT software to produce graphs and analyses tailored to the way you work.

SAS/INSIGHT software may be of interest to users of SAS/Genetics software for interactive graphical viewing of data, editing data, exploratory data analysis, and checking distributional assumptions.

SAS/STAT Software

SAS/STAT software includes procedures for a wide range of statistical methodologies including

- logistic and linear regression
- censored regression
- principal component analysis
- variance component analysis
- cluster analysis
- contingency table analysis
- categorical data analysis: log-linear and conditional logistic models
- general linear models
- linear and nonlinear mixed models
- generalized linear models
- multiple hypothesis testing

SAS/STAT software is of interest to users of SAS/Genetics software because many statistical methods for analyzing genetics data not included in SAS/Genetics software are provided in SAS/STAT software.