BIOSTATISTICS TEACHING IN THE NEW GENOME ERA

Júlia Pavan Soler
University of São Paulo, Brazil
Alexandre Pereira
University of São Paulo Medical School, Brazil
pavan@ime.usp.br

The objective of this article is to introduce a multidisciplinary biostatistics course program, targeting professionals and academics involved in researching genetics and genomics. The program is designed to adapt statistical concepts, uses and language to this field, which is at the forefront of knowledge. The idea underlying this initiative arose through a research project conducted by physicians and statisticians of the University of São Paulo, Brazil, whose purpose was to identify genetic determinants associated with cardiovascular risk factors in the Brazilian population.

INTRODUCTION

The application of statistical methods to genetics has a long history, starting with Fisher's contributions in the thirties and extending to the recent achievements of genome projects. This interchange of knowledge has been regarded as crucial in ensuring the progress of molecular biology and in the radical changes that have been occurring in medical practice as a result of new methods of disease diagnosis, treatment and prevention.

Recent progress in genomics have allowed us to identify genetic components involved in the complex mechanism of cell regulation of many of the diseases that represent some of the greatest healthcare problems in the world (such as diabetes, hypertension, obesity, depression, alcoholism, cancer, etc.). Despite the optimism justified by the success of several studies, the genetic architecture of such diseases has a multifactorial nature, involving multiple genetic and environmental components, and the analysis of how the genes operate and interact among themselves and with the environment continues to pose a major challenge for scientists. It is generally believed that multidisciplinary initiatives involving the efforts of different medical, statistical and bioinformatics specialists are crucial for the success of these research projects.

Within this set of circumstances, the teaching of statistics to physicians and biostatisticians has undergone its own paradigm changes, evolving from a standardized biostatistics course to a more targeted multidisciplinary program that adapts concepts, uses and language to this field, which stands at the forefront of knowledge. The key issue is to introduce the main sources of genetic data and methodologies for analyzing these data, for the purposes of genetic mapping and an understanding of interconnected biological systems. Although these subjects differ little from what is commonly taught in standard biostatistics programs, some particularities arise, such as the formulation of the genetic components of the quantitative variation and of the possible measures of association between genes.

In this article, we present a proposal for a biostatistics course program for academics and professionals involved with research in the genomic field. The proposal is concerned with employing teaching strategies capable of making the theme easy to understand, alternating practical lessons, which encourage active understanding, with lectures, centered on consolidating concepts. The level of detail and degree of formalism with each topic is dealt theoretically is conditional upon the student's statistical background.

MOTIVATION

In Brazil, cardiovascular diseases are the chief cause of death, accounting for 30% of the deaths in the country (The Brazilian Institute of Geography and Statistics – IBGE, 2000). This class of diseases is considered complex because environmental and genetic factors, as well as their interactions, affect its etiology and control. In mid 2003, medical, statistical and bioinformatics researchers from the University of São Paulo, Brazil, as well as from other research centers met to put into practice an innovative project in the country, whose objective was to identify and characterize cardiovascular risk factors among the Brazilian population, including

those of a genetic origin. The first experiences of this multidisciplinary project indicated a need for establishing a common language among the members of the group, as a means of setting a common ground for discussion and reaching a consensus on the experimental plane of the genetic data collection. In order to provide a response to shortcomings of this nature, we invested in proposing a biostatistics program that emphasizes experiment planning and genetic data analysis.

A BIOSTATISTICS PROGRAM IN THE NEW GENOMIC ERA

Statistical genetics is a fairly vast field of knowledge that justifies, per se, the development of a specific discipline program. Different student or even professional profiles seek courses of this type, driven by their need, to varying degrees, to acquire a more solid base in the fundaments of genetics, statistics or both. The book by Elston *et al*. (2002), for instance, is a resource that covers important fundaments of both genetics and statistics. On the other hand, for those who are more interested in a deeper statistical formulation of genetic issues, the book by Sorensen and Gianola (2002, p. 740) is an alternative. We describe below a Biostatistics course program in which the specific genomics content is incorporated into topics commonly covered in standard courses.

1. Classification of Variables (Phenotypes and Genotypes)
Practical Lesson: Visualization of Quantitative Variation (Lab 1 activity described in the next section)
Lectures: The Central Dogma of Biology (from the DNA to phenotypes), examples of Mendelian phenotypes (categorical variables controlled by only one gene) and complex phenotypes (quantitative variables controlled by many genes, environmental factors and interaction effects), random variables, probability models

2. Design of Experiments
Practical Lesson 2.1: Inbreeding Experiments (controlled crossings designs) - the goal of the studies, F2 design, backcross design, databases, disturbing biological systems
Practical Lesson 2.2: Experiments using natural populations - the goal of the studies, pedigree data sets, molecular marker maps, exploring familial data sets, genotype and phenotype data sets
Practical Lesson 2.3: Association Studies - the goal, case-control design, case and parental controls design
Practical Lesson 2.4: Microarray Experiments - gene expression data, the control of sources of variability, data normalization. In this lesson, there is use of computer resources, such as some of the Bioconductor functions that are available on http://www.bioconductor.org.

3. Probability Models
Practical Lesson: Linkage analysis (Lab 2 activity described in the next section)
Lectures: Binomial model, normal model, mixture distributions, linkage between genes, genetic equilibrium laws (Hardy-Weinberg and linkage disequilibrium), association between genes

4. Statistical Inference
Practical Lesson: Recombination fraction, Lod Score statistics, cytogenetic distance
Lectures: Likelihood function for different design of genetic experiments, genetic parameters (allele probability, genotype probability, recombination fraction, proportion of identical alleles by descent), Lod Score statistics

5. ANOVA Models
Practical Lesson 5.1: Factorial designs 3^k (k genes, 3 levels) - estimates of genetic effects and heritability measures
Practical Lesson 5.2: Multifactorial disturbance of biological systems
6. Regression Models

Lecture: Interval regression models for gene mapping in inbred experiments - specifying genetic effects as fixed components in the model, epistasis and pleiotropic effects
Practical Lesson: The use of specific computer programs for gene mapping, such as Cartographer (http://statgen.ncsu.edu/qtlcart/WQTLCart.htm)

7. Variance Component Models
Lecture: Variance component models for gene mapping in pedigree experiments: specifying genetic effects as random components in the model, epistasis and pleiotropic variances
Practical Lesson: The use of specific computer programs for gene mapping, such as *Solar* (http://www.sfbr.org/public/software/solar/index.html)

8. Contingency Tables Analysis
Practical Lesson: The Simpson Paradox in genetic association studies
Lectures: Case-control design, cases and parental controls, TDT tests, population stratification and the problem of false positives, confounding effects due to genetic admixture

9. Building Statistical Models for Gene Mapping
An activity that focuses on generating and analyzing genetic data in association with a specific set of research circumstances

The level of detail of each item may depend on the available number of course hours, the students' statistical background and on the particular interests of the student group. For instance, consider a student who has a basic knowledge of ANOVA and regression models. How can one introduce more elaborate models, such as models of variance components? We have used the strategy of initially exploring and consolidating the knowledge that the student has already acquired, emphasizing the non-random, i.e., the fixed nature of the predictive variables (a gene, for example), which model the expected mean of the response variable (a phenotype of interest); then, we illustrate the usefulness of this type of statistical modeling for the problem of mapping genes in controlled crossings designs, where it is natural to assume that all the levels of the genetic factor are a known element. After that, the problem of genetic mapping in natural populations is introduced, where the crossings are random and there is such genetic heterogeneity that the sample should only contain a few of the possible genes that control the phenotype in the population, i.e., a genuinely random genetic factor. In this case, limitations for using regression models are discussed and, as the students are highly motivated, the construction of a more general statistical model, in which the effect of the gene is assumed to be random, is conducted in a natural way. Although some mathematical details may be required, such as the derivation of the expressions for the mean and phenotypical variance, we emphasize more the method for building the model and the interpretations that result from it than its theoretical properties, such as the statistical inference procedures involved. In this type of course, by teaching the students to interpret things, we train them to think about the concepts and ideas associated to these things.

To illustrate this knowledge exchange a little more, other topics have been touched upon in the course by taking examples from Genetics, such as: (i) to show how different observational studies in Genetic Epidemiology avoid the problems of confounding variables by using family data; (ii) to explore microarrays experiments in order to introduce parametric and non-parametric normalization methods for data that aim at controlling sources of external variation via statistical adjustments.

PRACTICAL LESSONS

Below, we briefly describe two practical lessons among the several that the program suggests. The material of the other lessons may be requested through our e-mails.

Lab 1: Visualization of Quantitative Variation

Generate the phenotypes *Small* and *Tall* described below and present the histogram of the resulting distribution.

Phenotype *Small* is influenced by six dominant diallelic genetic loci, whose allelic frequencies within and between the loci are in equilibrium. Suppose, furthermore, that there is no impact from the environment and that the six loci operate in a cumulative way (without interaction effects). Consider that the genotype states, QQ and Qq, of each one of these loci add the value of one unit to the phenotype value and that $P(QQ \cup Qq)=0.75$.

Phenotype *Tall* follows the same regulation model as phenotype *Small*, but it is controlled by 24 genes.

Figure 1 shows the histogram of a simulation of both phenotypes based on the binomial probability model. An additional motivation can be obtained by extending the simulation of quantitative phenotypes exercise to more real situations, involving mixture distribution models as considered by Jansen (2003).

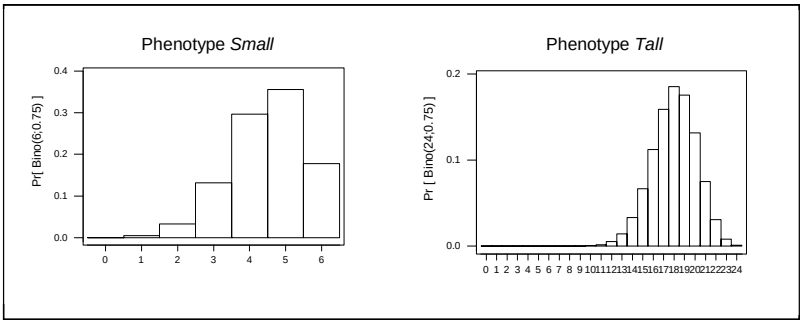

Figure 1: Distribution of quantitative phenotypes

In practical lesson *Lab 1*, statistical themes are introduced, such as the random experiment, sample space, random variable, sources of variation, predictor variables, probabilistic models, simulation, observed and expected frequencies, distribution of frequencies and also genetic concepts, like Mendelian and quantitative phenotypes, decomposition of quantitative phenotype into numerical functions, mixture distribution into quantitative phenotype understanding, and the various ways of genetic inheritance.

Lab 2: Genetic Linkage

Consider that molecular marker loci N and L will be used to inform the (unknown) location of a gene involved in the regulation of a recessive Mendelian disease. Note that under this segregation model, one can deduce that the individuals who have the disease carry two homologous copies of the allele responsible for the condition, i.e., DD, and that the unaffected parents with affected children must carry the Dd genotype. Furthermore, assume that the genotype of the N and L markers in the parent generation is known and follows the standard shown on Figure 2. Finally, assume that the corresponding genotypes of loci N and L in the children generation are known, being given by the following random process:

To determine the *genotype of locus N* in the children generation, throw a die to find out which allele was transmitted by the mother and which one was transmitted by the father. If an even number faces upwards (2, 4 or 6), the allele transmitted was N2 and N4, respectively. Otherwise, the allele transmitted was N1 and N3.

To determine the *genotype of locus L* in the children generation, throw the die again. If any of the numbers 1, 2, 3, 4 or 5 is facing up, the alleles transmitted by the father and by the mother were N1 and N3, respectively. Otherwise (if 6 faces upwards), alleles N2 and N4 were transmitted.

Perform the experiment described above ten times, i.e., consider ten families with the same structure indicated on Figure 2. Make a table with the results of the genotype data obtained and, for each one of the children, in each marker locus, calculate: the number of recombinations and the number of alleles that are identical by descent (ibd) shared between the siblings. One realization of this experiment is shown on Table 1. Based on these results, answer the following: does the distribution of the number of recombinations (of the number of ibd alleles) follow the

same pattern in the two loci, N and L? How can these results be used for the purpose of genetic mapping?

In practical lesson *Lab 2*, two fundamental concepts in the genetic mapping of disease are introduced, the linkage (distance) between genetic loci and the number of alleles shared between individuals. The throw of die as we are proposing makes the empirical simulation of genotypes, in such a way that based on the sample results it is possible to define two important numeric characteristics, the number of recombinant alleles and the number of alleles that are identical by descent (ibd). From these variables we can understand the laws of the pattern of genetic variation, for example, that siblings share more alleles in linked loci than in non-linked loci (where it is expected that only 1 allele is shared). As a result of this motivation we can build a statistical test to determine if a marker locus is close to a disease regulatory locus, for example, by comparing the average number of ibd alleles in the marker locus with 1. In addition, for this statistical test we introduce the lod score metric commonly used in the Genetic literature.

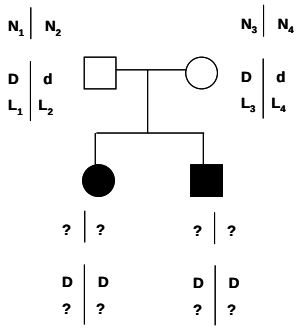

Figure 2: Pedigree and genotype data

Table 1: Executions of the random experiment

	Genotype		# of recombinations (# of ibd alleles)	
Family	Daughter	Son	Locus N - unlinkage	Locus L - linkage
1	(N1,N4) (L1,L3)	(N2,N4) (L1,L3)	3 (1)	0 (2)
2	(N2,N3) (L1,L3)	(N1,N3) (L2,L3)	1 (1)	1 (1)
3	(N2,N4) (L1,L4)	(N1,N4) (L2,L3)	3 (1)	2 (0)
4	(N2,N4) (L1,L4)	(N2,N3) (L1,L3)	3 (1)	1 (1)
5	(N1,N4) (L1,L4)	(N2,N3) (L1,L3)	2 (0)	1 (1)
6	(N2,N4) (L2,L3)	(N1,N3) (L1,L3)	2 (0)	1 (1)
7	(N2,N4) (L1,L4)	(N1,N3) (L1,L3)	2 (0)	1 (1)
8	(N2,N4) (L1,L3)	(N2,N3) (L1,L3)	3 (1)	0 (2)
9	(N2,N4) (L1,L3)	(N1,N3) (L1,L3)	2 (0)	0 (2)
10	(N2,N4) (L1,L4)	(N2,N4) (L1,L3)	4 (2)	1 (1)

DISCUSSION

In this article, we introduced a biostatistics course program designed for academics and professionals involved in genomic research. In an attempt to replace the “boring” adjective commonly ascribed to applied statistics courses by “pleasurable,” it is advisable to follow certain pedagogical recommendations (Moore, 2000; Garfield, 1995). In this context two fronts are planned for the program: to emphasize the statistical thought process that underlies the genetic data analysis methodologies and also the genetic concepts that are being statistically modeled.

We share the proposal defended by various authors (for example, Nolan and Speed, 1999), on the use of the case study and real examples when teaching Statistics. To this end the state of the art in Genetics and Genomics offers a promising scenario, in which highly motivated students, conscious of the need to use statistical methodologies (many times not trivial) and determined to keep up to date, when they take a course in Statistics based on their particular area of interest, respond much more quickly to the demands and spare no effort when it comes to obtaining the desired training in Statistics. Furthermore, our proposal in the Biostatistics course based on Genetics may be accomplished, either in full or in part, with on-going training programs in Statistics that aim at updating and improving statistical methodology as applied to relevant problems, such as the series of seminars presented by Deutsch (2002).

Additionally, we highlight that holding the practical lessons in a group should be ensured in programs of this type, because this makes it easier to introduce concepts that are often hard to grasp solely through a lecture. The introduction of specific statistical packages for analyzing large genetic data bases (for example, *Cartographer*, *Solar* and *Bioconductor*) and consulting web pages (for example, http://linkage.rockefeller.edu, http://www.genome.gov/education) are also crucial for programs of this type.

Student performance evaluation in the course was conducted through an activity centered on data generation and analysis, requiring that the students plan all the stages of a genomics experiment and build the statistical model for data analysis. Our experience in applying the said program is still in its infancy, in that the course has only been taught four times, to classes with an average of 15 post-graduate students, from both the statistics area, or otherwise, but all involved in some way or another in research in the area of Genetics. It is worth highlighting that this latter criterion has been the only demand for the student to be accepted to the course. However, the results (not shown here) have pointed to better student performance when the student has some prior knowledge of statistics. This may be an indication of the need to establish for future courses the pre-requisite of at least one full semester in Statistics, in order to have more guarantee that the benefit being targeted is achieved.

Accomplishment of the proposed course program is flexible, given that the level of detail with which each topic is dealt may be decided based on the students' background in each of the areas involved. The problem should be analyzed in different dimensions. While standardized "adjustments" of language uses and basic concepts is a need for both researchers in the biological and statistical areas, one can not underestimate the need for a deeper introduction and explanation of the statistical formulation of such items.

Our view, given the radical advances in medical practice where methods of disease diagnosis, treatment and prevention are concerned, is that there is a conspicuous need for biostatistics courses to take into account genome-related topics, thereby filling a gap in professional education in the areas of medicine, statistics and bioinformatics, all of which are increasingly involved with research in this new field that stands at the forefront of knowledge.

ACKNOWLEDGMENTS

This study had the financial support of FAPESP the São Paulo State Foundation for Research Support (*Fundação de Amparo à Pesquisa do Estado de São Paulo*), SP, Brazil (FAPESP Grant 01/00009-0). Physical and data processing resources were guaranteed by IME – USP, the University of São Paulo Mathematics and Statistics Institute (*Instituto de Matemática e Estatística da Universidade de São Paulo*), SP, Brazil.

REFERENCES

Deutsch, R. (2002). A seminar series in applied biostatistics for clinical research fellows, faculty and staff. *Statistics in Medicine*, 21, 801-810.

Elston, R. C., Olson, J. M., and Palmer, L. (2002). *Biostatistical Genetics and Genetic Epidemiology*. Chichester: Wiley.

Garfield, J. (1995). How students learn statistics. *International Statistical Review*, 63(1), 25-34.

Jansen, R. C. (2003). Quantitative trait loci in inbred lines. In D. J. Balding, M. Bishop and C. Cannings (Eds.), *Handbook of Statistical Genetics* (second edition), Vol 1, (pp. 445-476). Chichester: Wiley.

Moore, T. (Ed.) (2000). *Teaching Resources for Undergraduate Statistics*. Washington: Mathematical Association of America.

Nolan, D. and Speed, T. P. (1999). Teaching statistics theory through applications. *The American Statistician*, 53(4), 370-375.

Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC methods in Quantitative Genetics*, 2nd printing. New York: Springer-Verlag.