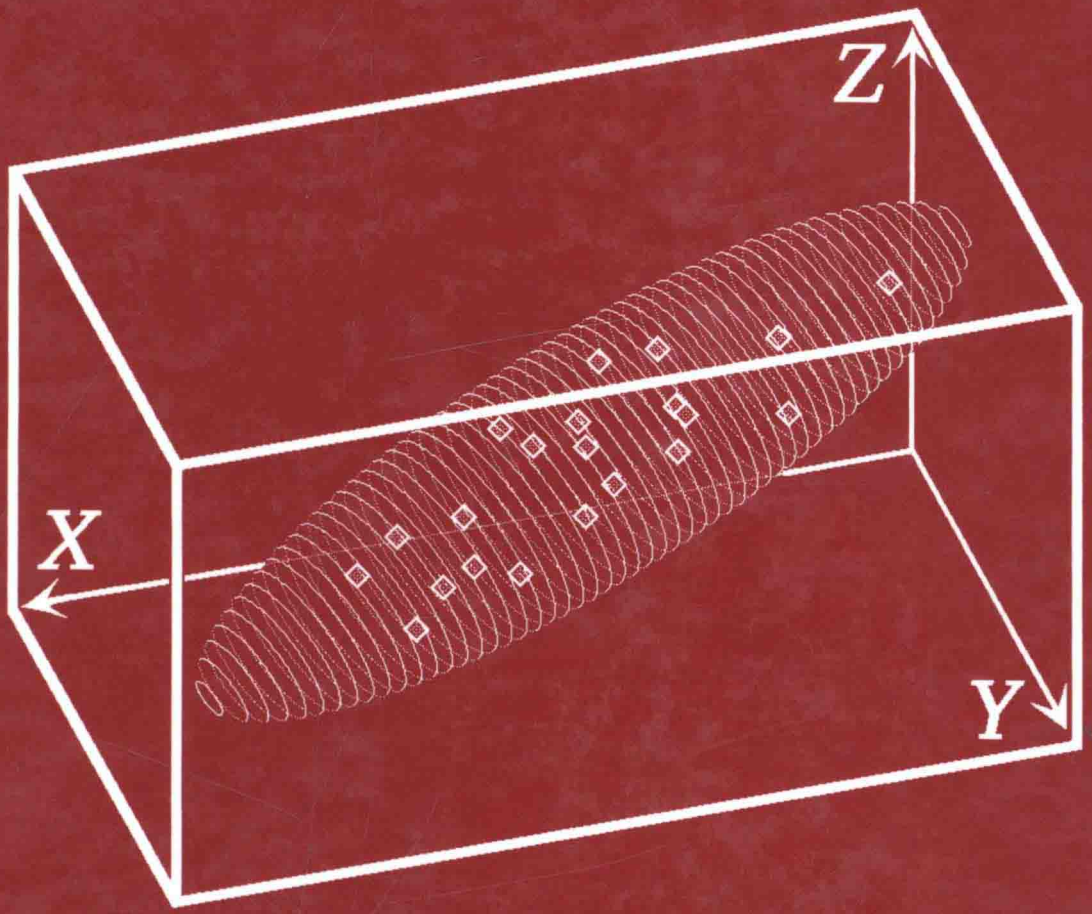
INTRODUCTION TO BIOMETRY



PIERRE JOLICOEUR

Introduction to Biometry

Pierre Jolicoeur
Department of Biological Science
University of Montreal

Forewords by

Richard Tomassone
Department of Mathematics and Computer Science
National Agronomical Institute, Paris

and

Jean-Marie Legay
Claude-Bernard Lyon I University, Villeurbanne

Kluwer Academic / Plenum Publishers
New York, Boston, Dordrecht, London, Moscow

This book has been translated by the author, with assistance from his wife, Veronika Meinow, from the fifth edition *Introduction à la biométrie*, which will be published in 1999 by Décarie Éditeur, Inc.

The front cover illustration is based on an original analysis of data published in 1905 by J. S. Haldane and J. G. Priestley, who pioneered the study of human respiration.

ISBN 0-306-46163-3

10 9 8 7 6 5 4 3 2 1

A C.I.P. record for this book is available from the Library of Congress

Printed in the United States of America

Introduction to Biometry

Foreword

The collaboration between biologists and mathematicians, mentioned by Jean-Marie Legay in his foreword to the first French edition of Pierre Jolicoeur's book, is of course as necessary as ever. The interactions between the two different fields of thinking must go on! The mathematician, on the one hand, may perhaps more easily function within his own universe without suffering too much from a lack of collaboration, although he will be intellectually frustrated to see that the tools he conceives are not readily used by other scientists, as he thinks they should be. The biologist, on the other hand, can hardly ignore mathematical and statistical methods: he must progress in his own field, developing his ideas and making decisions with tools often created by others but indispensable in order to extract all information available in experimental data.

Language is undoubtedly of prime importance to transmit knowledge and motivations from the mathematician to the biologist and vice versa, a fact too often neglected by many scientists. In order for the objectives of a study to be well understood, and for mathematical formulae to be applied correctly to life phenomena, communication through well-chosen language is as important as technical training, but this is often underestimated in teaching, especially at the university level. Biometry aims at being the melting pot of mathematical statistics and biology, which makes an aptitude for dialogue mandatory in the biometrician. This aptitude can only be developed through an intimate acquaintance with both mathematical (statistical) tools and a particular field of application, as illustrated by the case of Pierre Jolicoeur.

The author of this textbook belongs to a group of scientists who feel a strong yearning to fill this need. His treatment of the subject is simple and clear, yet highly rigorous: it should satisfy all kinds of readers. The mathematician will find the essentials of what must be learned and taught, and undoubtedly also a manner of presenting mathematical notions in such a way as to adapt them to biological contexts. The biologist will get even more. First, he will obtain tools which can be applied immediately to his own problems, something which he might consider as a strictly decent minimum... In addition to basic statistical formulae, however, the biologist will also discover a simple and intelligent manner of reasoning which can be used to adapt statistical models to complex biological situations. He will realize that biometry enables him to emphasize the salient features of an analysis while discarding the random peculiarities which are present in any experiment or observation. But many other categories of potential users, besides biologists, could profit from this book, and all those whose results are affected by random should be able to transpose to their own field the ideas and techniques discussed here.

Admittedly, contemporaneous biometry is a rapidly evolving science (or is it a craft?), and the new tools which are created every day are not all mentioned in this textbook, but this was not the author's purpose. However, once the reader has read the book and understood its contents, and has applied some of the methods to his own data, he will have reached a "point of no return" in his growing acquaintance with biometry. The biometrical way of thinking will have become anchored in his mind, and he will be able to read and understand more specialized and possibly more technical textbooks, if necessary. But he will already have acquired the essential basis needed to develop his own personal reflections.

I think that the readers of this book should be aware of their chance. As for myself, I can only regret not to have had Pierre Jolicoeur's book at hand when I began my career as a biometrician!

Richard Tomassone

Foreword to the first French edition (1991)

The necessity of multidisciplinary studies is widely acknowledged today. The training of mathematicians should therefore include some biology but, except in a few universities or other schools of higher learning, this is still generally not done on a satisfactory scale. Similarly, experimenters, including biologists, can no longer get along without some knowledge of mathematics, which they will need either in data analysis or in model building. Whether experimental scientists carry out such activities themselves or collaborate with mathematicians, they must understand quantitative techniques, and be aware of their possibilities and limitations.

In the various biological sciences, including medicine and agriculture, many students are therefore faced with a major problem since, although they do not wish to take advanced training in mathematics, their studies and their own research will require them to know and understand some mathematical concepts and techniques.

Fortunately, some biologists and some mathematicians have investigated questions involving biology and mathematics for decades (models of differential growth, population dynamics, spatial dispersion, prey-predator interactions, etc.). Their joint activities have given birth to an interface field: biometry. Not surprisingly, courses in biological mathematics or in mathematical biology are often taught by biometricians, who do research at the frontier of both fields.

As a biometrician, Pierre Jolicoeur is known throughout the world for his early studies of complex biological phenomena. He pioneered the use of multivariate statistical methods in morphometrics, in the case of bilateral symmetry for instance. More recently, he has developed original nonlinear models for somatic growth. Pierre Jolicoeur's interdisciplinary stand is clearly shown by the title of this book. Drawing upon his experience both as a research worker in biometry (see the list of his publications) and as a teacher, the author is undertaking the difficult task of making statistics palatable to biologists and biology students.

What makes this textbook markedly original is that elementary notions of probability and statistics are presented in an unabashedly simple and clear fashion, with helpful comments, and are illustrated using new and often personal data sets. Moreover, the text includes frequent hints at more elaborate developments, which makes careful reading very rewarding.

While the reader is generally spared detailed rigorous mathematical demonstrations, assumptions and hypotheses are always clearly stated, and the advantages and disadvantages of the various methods and approaches with respect to the aims pursued are always discussed. The author has evidently chosen to guide the reader's efforts toward a qualitative understanding of procedures rather than the technical derivation of already-known results.

Emphasis is placed on the distribution concept, which appears generally adequate for the description of biological variation. After presenting the statistical distributions most frequently useful in biology, the author discusses their utilization for testing hypotheses concerning means, variances, frequency tables, goodness of fit, and simple or partial and multiple correlations and regressions. While the total number of variates is limited to two or three[1] in order to make it easy for the (beginning) reader to visualize relationships, natural extensions are evoked whenever possible.

While the author generally does not treat his subject in a highly technical manner, his attempt to give the nonmathematical reader a deep and intuitive understanding of biological statistics is obviously a considerable challenge. Moreover, by letting beginning students of biometry glimpse at more advanced topics, this book may stimulate some of them to pursue their studies further. Let us wish the author all the success he deserves.

Jean-Marie Legay

[1] Editor's note: in the present edition, the simultaneous analysis of more than three variates is covered in chapters 24, 25 and 29 to 34.

Contents

The statistical tables most frequently used in biometry

Introduction

The word *biometry* comes from the Greek (βιος, *life* + μετρον, *measurement*) and means literally the quantitative study of life phenomena. Since living organisms generally differ from each other in size and shape as well as in their functions, the study of these differences requires statistical methods. Consequently, the word *biometry* is often considered as a synonym of *biostatistics*. However, biometricians generally pay more attention to biological aspects than statisticians would do. For many years, it has been realized that a knowledge of biometry is a must for practicing biologists as well as for graduate students. Therefore, most universities are now aware of the necessity of giving their biology students at least one introductory course in biometry.

The present textbook is aimed at university level biology students as well as at biologists wishing to improve their knowledge and understanding of biometry. The author's viewpoint is intermediate between *classical statistics* and *data analysis*. In addition to presenting a broad spectrum of statistical methods, the author has emphasized understanding as much as possible, in order to enable readers to become gradually self-reliant. However, since this book will presumably be read mostly by biologists, in some cases technical explanations are given in ordinary English rather than through formal mathematical demonstrations.

The coverage of the so-called nonparametric methods is limited to the analysis of frequency tables (chapter 15), tests of goodness of fit (chapter 16), and the sign test (section 17.11): the author believes that parametric methods provide better illustrations of the logical role which statistical methods can play in scientific research, the statistical parameters corresponding to the desired theoretical knowledge. Moreover, nonparametric methods somewhat duplicate their parametric counterparts and are redundant to some extent in this respect. Finally, nonparametric techniques are generally less sensitive ("powerful") than their parametric equivalents. The frequency distribution of biological data is often similar enough to the normal distribution (chapters 5, 19, 23, 29), or can be made similar enough through transformations, for methods based on the normal distribution to be used. When other nonparametric methods are truly needed, the reader should consult Sprent's thoughtful introduction (1993).

Biologists and biology students are more strongly motivated for the study of statistical methods when the latter are illustrated on genuine biological data. For this reason, most examples in this textbook are based on real data extracted from scientific publications. In the few cases where artificial data have been used, they have been made to simulate biological reality as closely as possible.

An effort has been made to organize the subject matter in the most efficient pedagogical order. Whenever there is overlap between chapters or between the sections into which they are subdivided, cross-references are given. Exercises (problems) are not included in the present edition but may be published separately in the future.

The five statistical tables which are most frequently used in biometry have been entirely recomputed using algorithms discussed by Abramowitz and Stegun (1968) and by Kennedy and Gentle (1980). These tables cover a particularly extensive range of numbers of degrees of freedom in order to make interpolation seldom necessary.

Tabulated values generally include four or five digits, which should all be accurate since computations were carried out in triple precision. Moreover, the risk of accidental errors has been eliminated by transferring numerical entries electronically. In order to reduce ambiguities and to promote methodical working habits, all tables are given in terms of the *cumulative probability*, i. e. the probability that the numerical value of a *variate* (random variable) is less than the tabulated value.

Some readers may wonder whether it is still pertinent to provide extensive statistical tables at a time when personal computers enable anyone to compute almost instantly the probability of an observed value or, inversely, the value corresponding to a specified probability. The decision to include such tables in this textbook was made for pedagogical reasons: the author is convinced that a university student's familiarity with statistical methods would remain superficial if his training were limited to using computer programs developed by other persons. Nevertheless, every biometrician should possess and use either a good scientific calculator or a personal computer.

The author hopes that this textbook will help the reader to become familiar with the basic notions of biometry and to discover the interesting challenge of applying mathematics and statistics to biology. Should any reader notice obscurities or errors, he is invited to contact the author, who would be grateful for constructive suggestions. Thanks are expressed to the readers of the first three French editions who suggested improvements.

Several generations of students incited me, by their questions and comments, to present many topics more clearly. Several research workers, as indicated within the text, gave permission to base examples on their interesting data. Many years ago, Dr. James E. Mosimann, then at the Department of Biological Sciences of the University of Montreal, stimulated my early interest in biometry, and Professor Stanley W. Nash, of the Department of Mathematics of the University of British Columbia, answered many questions when I began to develop an acquaintance with multivariate analysis. Professor Jean-Marie Legay, of Claude-Bernard Lyon I University, France, and Professor Richard Tomassone, of the Department of Mathematics and Computer Science of the National Agronomical Institute, Paris, kindly accepted to write forewords to the first French edition (1991) and to the present edition. Professor Jacques Pontier, also of Claude-Bernard University, made valuable suggestions. Professors William H. Kruskal and Stephen M. Stigler, of the Department of Statistics of the University of Chicago, provided historical information. My children, my wife, Veronika Meinow, Mr. André Décarie, Mrs. Edenise Garcia and Mrs. Anne-Marie Blais contributed to the improvement of the text and the illustrations. Finally, my wife helped checking the English translation. I sincerely thank all persons mentioned here, as well as many others whose names are not listed for conciseness.

I dedicate this book to the memory of my parents and to the happiness and success of my children, Lucie, Francine, and André.

Montreal, July 1, 1998 Pierre Jolicoeur

Author's address:

Département de Sciences biologiques, Université de Montréal, Case Postale 6128, Succursale Centre-Ville, Montréal, Québec H3C 3J7

or

1226, Rang Égypte, Case Postale 160, St-Valérien, Québec J0H 2B0

Chapter 1
Looking at quantitative biological data through scatter diagrams

Section 1.1: scatter diagrams

In applied statistics in general, and in biometry in particular, graphical representations and numerical techniques are complementary and often equally important. One of the most useful graphical methods in statistics is the *scatter diagram,* in which variable measurements are represented simply by dots dispersed on a surface overlaid by Cartesian coordinate axes (named after the French mathematician René Descartes, 1596-1650). Such coordinate axes are usually perpendicular to each other. While a scatter diagram may have a single coordinate axis (one-dimensional dispersion) or more than two coordinate axes (three-dimensional or multidimensional dispersion), the most commonly used version is the *bivariate scatter diagram,* which contains two coordinate axes (two-dimensional dispersion) and where dots represent pairs of measurements (see figures 1.2.1 and 1.2.2).

Section 1.2: an example of the graphical examination of quantitative data

The numerical measurements in table 1.2.1 were obtained on undissociated human skeletons following archeological excavations in England, and are represented in a bivariate scatter diagram (figure 1.2.1, next page). Data given on any single line, in table 1.2.1, correspond to measurements made on the skeleton of a single individual.

Table 1.2.1

Length measurements of the left humerus and right humerus of female human skeletons; data drawn from the study of Münter (1936)

(data reproduced with the permission of Oxford University Press on behalf of the Biometrika Trustees)

Number of skeleton (subscript)	Left humerus (mm)	Right humerus (mm)
1	$X_1 = 311$	$Y_1 = 315$
2	$X_2 = 302$	$Y_2 = 306$
3	$X_3 = 301$	$Y_3 = 311$
4	$X_4 = 322$	$Y_4 = 333$
5	$X_5 = 312$	$Y_5 = 316$
6	$X_6 = 285$	$Y_6 = 292$
7	$X_7 = 305$	$Y_7 = 308$
8	$X_8 = 310$	$Y_8 = 318$
9	$X_9 = 328$	$Y_9 = 326$
10	$X_{10} = 304$	$Y_{10} = 309$

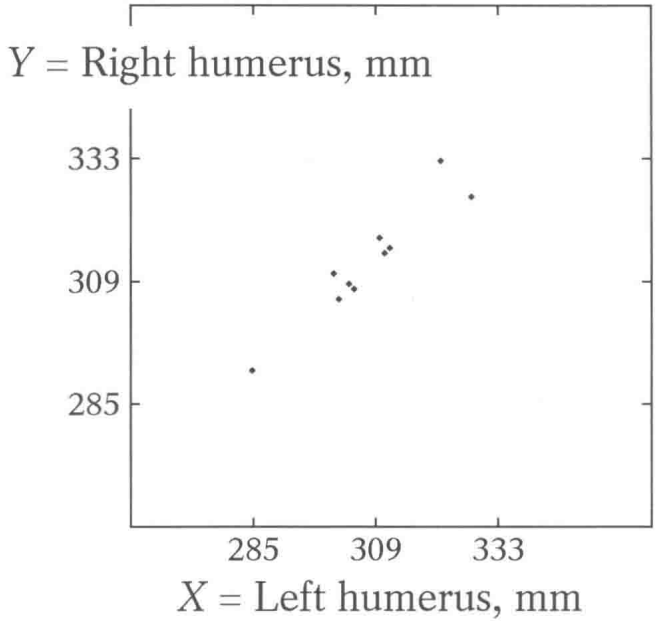


Figure 1.2.1
Scatter diagram of left humerus length X and right humerus length Y of ten female human skeletons; data from Münter (1936)

Similarly, in figure 1.2.1, each dot represents one individual (or, more exactly, its skeleton) whose left humerus length and right humerus length correspond to the coordinates on the abscissa (the X-axis) and on the ordinate (the Y-axis) respectively. A careful examination of this scatter diagram reveals one important feature of this small data set: dots are spread approximately along a hypothetical straight line going from the lower left corner to the upper right corner of the diagram. This suggests that a person having a particularly long left humerus tends to have also a particularly long right humerus; inversely, a person having a short left humerus also has a short right humerus. This diagram thus appears to show that, in the human skeleton, the left humerus and the right humerus tend to have closely similar lengths. In fact, this is an instance of the phenomenon of *bilateral symmetry,* which occurs in man, in many other animals, and even in certain structures of some plants.

If bilateral symmetry were perfect, one would expect all dots to lie exactly on a straight line passing through points having equal X- and Y-coordinates, of which the equation would be $Y = X$. Such a line has been drawn in figure 1.2.2. Bilateral symmetry is obviously not perfect, since dots do not all lie exactly on the line $Y = X$. Moreover, 9 out of the 10 individual dots, as well as the mean dot $(\overline{X}, \overline{Y})$ (see chapter 4), lie above the line $Y = X$, in the part of the scatter diagram where $Y > X$, that is where the right humerus is longer than the left humerus. As numerical methods will confirm in sections 9.6, 9.9 and 17.11, the right humerus thus seems slightly longer that the left humerus in most female human skeletons.

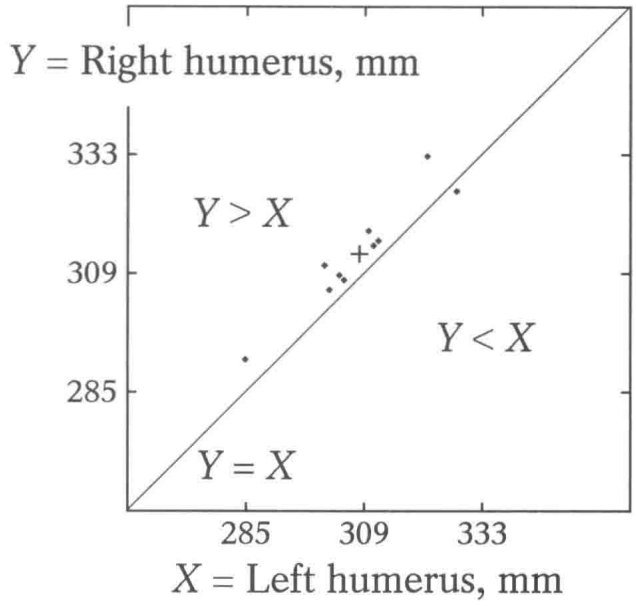


Figure 1.2.2
Scatter diagram of left humerus length X and right humerus length Y of ten female human skeletons; the diagonal line $Y = X$ represents the hypothesis of bilateral symmetry and the cross (+) represents the mean dot $(\overline{X}, \overline{Y})$

One may wonder whether this slightly greater length of the right humerus might be related to the fact that, in all contemporaneous human populations, and even in earlier populations according to the analysis of prehistoric drawings (Coren and Porac, 1977), most individuals (from 90% to 95%) are right-handed. It is rather striking that the simple graphical examination of such a small data set should provide such interesting information and raise such fascinating questions.

As for bilateral symmetry, the functional asymmetry of the human brain has received considerable attention from neurobiologists during the last two decades, the question being reviewed by Bradshaw and Rogers in 1993. In a multivariate statistical analysis of limb bone asymmetry in man and in the North American Marten, Jolicoeur (1963c) has discussed the hypothesis that, in animals possessing a cephalic pole (a head), bilateral symmetry (with respect to a median plane) is an adaptation to rapid straight-line locomotion in an environment vertically stratified because of gravity. Contrariwise, *radial symmetry* (with respect to a single axis) often occurs in animals whose cephalo-caudal axis is weakly differentiated and generally vertical, and which are fixed or move slowly without a marked directional preference, like Coelenterates (sea anemones and jellyfish) and adult stages of many Echinoderms (sea urchins and sea stars). In plants, *radial symmetry* is most frequent, but bilateral symmetry does occur in some diatoms and in the so-called *zygomorphous* flowers of several angiosperms (orchids in particular), which have a slanting habit and of which many interact strongly with pollinizing insects.

Chapter 2
Samples and populations, estimates and parameters

Section 2.1: samples and populations

The words *sample* and *population* have different and more restricted meanings in statistics and in biometry than in common everyday language. In statistics, a *population* (also called a *universe*) is a set of beings (or the set of qualitative or quantitative observations which can be made on those beings) about which information is desired and which is studied through a subset, called a *sample*, drawn from it at *random,* that is in as impartial (unbiased) a manner as possible.

In the biological sciences, but not in statistics, the word population implies the existence of living organisms differing from each other with respect to age and sex and able to reproduce themselves sexually or asexually. Biometricians must therefore always watch out and carefully distinguish the cases where the word population is used in its biological as opposed to its statistical meaning. However, this does not rule out the possibility that a particular group of living organisms may be justifiably considered as a *biological population* by a biologist and as a *statistical population* by a statistician or a biometrician.

As for the word sample, its statistical meaning is that of a subset drawn from the statistical population in order to get information about it. This subset generally includes several beings or observations, and is only exceptionally reduced to a single unit. On the contrary, in everyday language, a sample is often a single specimen, like a piece of fabric in textile marketing, a piece of rock in mineralogy, or a small amount of blood or urine in medical biochemistry.

The number of beings or observations included in a statistical sample is known as the *sample size* and is often denoted by the symbol N. As for statistical populations, finite populations, containing limited numbers of beings or of possible observations, are occasionally considered (sometimes for didactic reasons, see section 9.5), but many populations are made up of so many units that they can be considered as practically infinite. Moreover, even if a population is thought to be finite, its size is seldom known exactly. Ecologists are frequently interested in estimating the size of finite natural populations.

A statistical population is a whole which one tries to know by studying one of its parts. The logical operation through which findings made on the part are extrapolated to the whole constitutes a *generalization* (an *inductive reasoning* or *inference*). Let us remember that the conclusion of an inductive inference is uncertain, except in the rather special case of *complete induction* (where the whole population is studied).

Would it not be simpler to study all members of the statistical population directly instead of taking a sample? – Theoretically yes, but this would not always be feasible, since some statistical populations are extremely large. Moreover, some biological studies require living organisms to be killed: in such cases, only part of the population is usually taken in order to prevent extermination. Finally, the direct study of a whole population