1

What This Book Is About

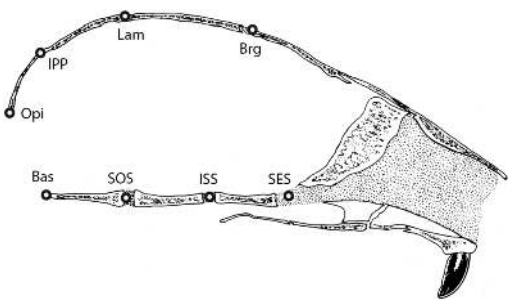
1.1 Overview: Where Morphometrics Draws Its Ideas

Morphometrics, the topic of this book, brings together the best modern methods for quantifying the patterns of variation of organismal form that we biologists are most interested in talking about. Sometimes the organisms under study are humans, in which case we usually refer to our work as an aspect of human biology or medicine. Or the research question may be classified under a rubric like zoology, botany, paleontology, or ecology. The objects of study, when they are not actually human subjects, might be living (in which case they may be domesticated, or farmed, or else studied in the wild), but sometimes they are dead or even extinct. They might be observed only once each, or over a short window of time corresponding to some physiological cycle, or perhaps over a substantial fraction of the life cycle or even all of it. Some investigators who make use of our morphometric tools are concerned with the evolutionary, developmental, genetic, or environmental processes that account for those patterns, while others pursue the consequences of those patterns for the biological functions they help govern (physiological energetics, locomotion, predation, reproduction, health), the time course of the organism's life, or the consequences for the ecological system(s) in which the organism can be found. Now a mature scientific interdiscipline, morphometrics combines knowledge bases and expert analytic strategies from geometry, statistics, and classical functional anatomy with the customary investigative styles and question constructions of zoology, paleobiology, medicine, and bioengineering in order to acquire and synthesize the information about organismal form that answers existing questions like those, or to pose new ones.

The underlying intellectual themes on which we draw for the tools and examples this book introduces hence range quite broadly. They are not the property of any single discipline, let alone the discoveries of any single

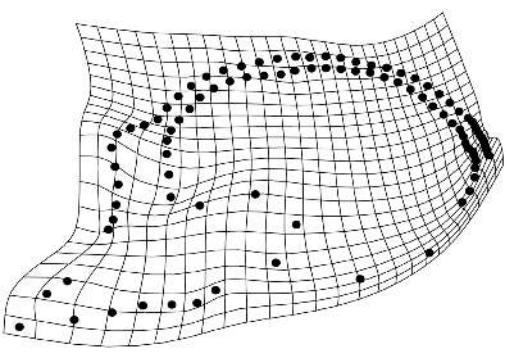
founding father (not even me) or small community of proud progenitors. Rather, they span nearly all the classic branches of the biological sciences and also nearly all of the allied disciplines that build tools for the bioscientist or supply the motivation or the language(s) used for research design, reporting, or dissemination. Here, anticipating nine of the later figures in this book, are some of those components:[1]

- From *comparative anatomy* comes the main thread that links our work to the classic language of organismal form and function:



  - ⊙ the **homologous anatomical structures** that help us keep track of our place – what we are pointing to – when we talk about variation of complex organismal forms.

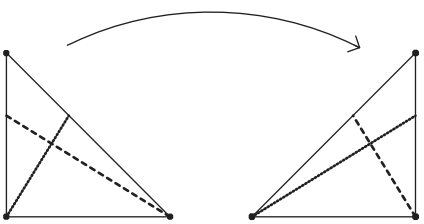
- From modern *cognitive psychology* comes our principal diagrammatic device:



[1] These copy or modify Figures 2.41, 5.2, 5.25, 2.5, 2.10, 2.12, 5.3, 5.55, and 5.83, respectively.

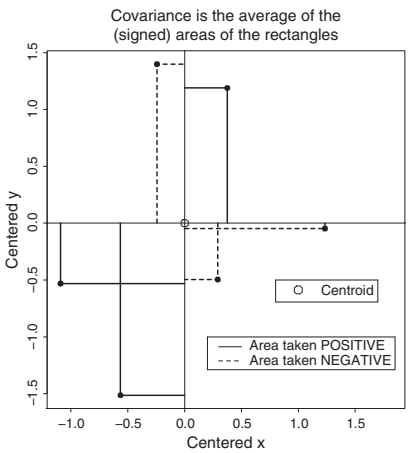
- ⊙ the **transformation grid** that we use to present an extended comparison of landmark configurations to the intelligent eye.

- From *mathematics* we borrow endlessly, including:
  - ⊙ the properties of **derivatives** that let us compute extrema of functions, such as sums of squares or the corresponding likelihoods, when we need those maxima or minima to summarize data;
  - ⊙ the theorems of **probability theory**, such as the binomial distributions that summarize tosses of coins, and the rules of **limits** that let us get from there to the bell curve;
  - ⊙ the geometry of **quadratic forms** and the ellipses, ellipsoids, parabolas, and the like that illustrate the theorems about them;

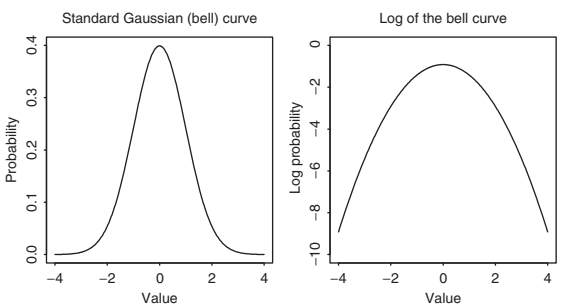


  - ⊙ the rules of **tensor algebra**, which authorize us to express shape changes of triangles as ratios along a pair of directions that remain at an invariant angle of 90°;
  - ⊙ the related rules of **matrix algebra**, which govern maneuvers such as inversion or eigenanalysis when applied to covariance structures; and
  - ⊙ the theorems about **function spaces** that justify the crucial insight that the thin-plate spline interpolant is the "smoothest possible map consistent with the given data," whatever those location data are.

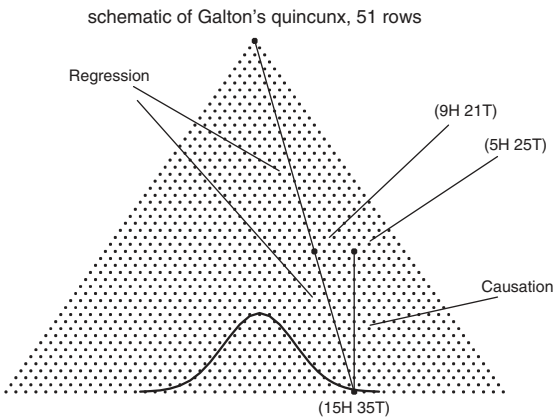
- From *statistics* come these ideas, among many others:
  - ⊙ the **least-squares** models that summarize varying empirical data via simple underlying geometrical structures such as lines or planes;
  - ⊙ the **covariance coefficient** that summarizes the relationship between two measured quantities – this is the numerator of the usual formula for the slope of the least-squares regression line;



- the **covariance matrices** that summarize the covariances of multiple measurements, all together, in one conceptual object susceptible to the mathematical manipulations of matrix algebra mentioned earlier;



- the Gaussian distribution or "**bell curve**," universal model for disorder in any natural-science system;
- the **Wishart distribution**, equivalent of the bell curve for assessing the disorder in a covariance matrix summarizing some empirical data involving more than one measurement;
- the **information matrix** and its inverse, which summarizes the sampling error variances and covariances of the parameters of any numerical model around their maximum-likelihood estimates;



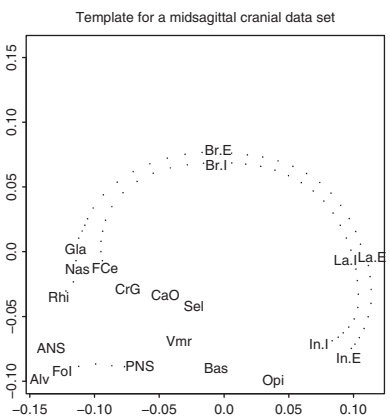
- ⊙ Francis Galton's **quincunx**, a machine that generates linear regressions as correlated bell curves – straight lines that arise even in the absence of any governing physical theory, and that account for regression to the mean as an intrinsic property of pure noise; and
- ⊙ the algebraic technique of **quadratic discriminant analysis** for classifying organisms among two or more groups that vary in their covariance matrices as well as their group averages.

- From *physics* come three crucial understandings of the disorder through which, under the guise of "variance," we must filter our biological explanations:
  - ⊙ the **Maxwell–Boltzmann distribution**, which accounts for the motion of gas molecules on the Gaussian model;
  - ⊙ the broader notion of **entropy**, in terms of which bell-curve disorder is the universal maximum, and the identification of information (whatever its origin) as the opposite of entropy; and
  - ⊙ the expectation that **mathematics is often unreasonably effective** in summarizing natural variability (as in, for instance, the actual bell-curve formula itself – why on earth should there be a π there?), once the measurement conditions are placed under sufficiently stringent experimental control.

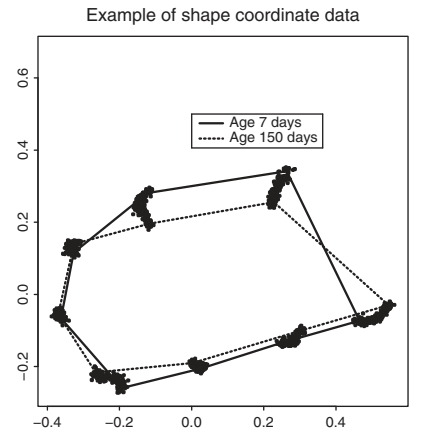
- From the *biological sciences,* broadly considered, come ideas such as

- ⊙ **morphological integration** (substantial but inconstant correlation among most pairs of measurements) as a ubiquitous property of numerical representations for every living system;
- ⊙ the centrality of **measures of extent** – lengths, areas, weight – and **measures of proportion** for summarizing patterns of growth or the relation of biological forms to their causes or effects;
- ⊙ the corresponding centrality of **path coefficients**, slopes of lines, for disentangling genetic or biophysical aspects of the underlying processes we are studying; and
- ⊙ **Brownian motion**, which identifies the disorder of particles in fluid suspension as tracked through a microscope with the disorder of gas molecules and, in a later realization, with the disorder that mathematicians learned to model under the heading of "random walk."

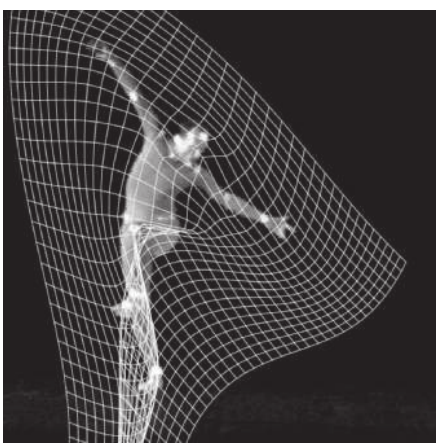
- Finally, from *morphometrics per se*, the literature of tools and expertise for numerical analyses of data on size and shape that this book surveys, we will be systematically reviewing such central strategies as



- ⊙ the protocols of **landmarks and semilandmarks** that allow us both to design good measurement schemes based on point, curve, or bounding surface locations and thereafter to report patterns of their differences or covariances by focus and extent;
- ⊙ the **shape coordinates** that convert these multiple point locations into measurement vectors for carefully formulated multivariate statistical analyses;



Shape coordinates for 144 rodent neurocranial octagons, Section 5.4



- the **thin-plate spline grids** that convert the customized multivariate statistical analyses of those vectors back into the coordinate system of the picture of the organism(s) the forms of which we are trying to explain; and
- the way that processes of growth or size allometry tend to suit a description by the first one or two **principal components** of shape coordinates almost as well as they do for multiple measures of extent (such as lengths).

The overall development of these ideas across this book falls into two halves with a link. In the first half, Chapters 2 and 3, the classic statistical technology of covariance-based least-squares linear modeling is reviewed as radically rephrased for morphometric studies of physical extents (quantities that add, like length, area, or weight). The second half, Chapter 5, builds on the technologies of the first to fabricate tools specialized for data that come as locations. We extend the covariance methods to incorporate a novel data type, the shape coordinates that convert the Cartesian location data into a format that the covariance-based methods can take hold of and then report using our preferred grid metaphor, the thin-plate spline. In-between these two thrusts is a bridging Chapter 4 that combines the several appearances of the linking statistical theme across the diverse twentieth-century literatures in which they originally appeared: themes as diverse as the Wishart distribution of covariance matrices, the information matrix, the singular-value decomposition, factor analysis, principal component analysis, and the relation of all these to the complexly structured measurement schemes that capture the variations of organismal form study by study. Outside of the morphometric toolkit itself, the shape coordinates and thin-plate splines, hardly any of these tools have the same inventor; it is their combination in dataflows that is crucial, from organismal anatomies through shape coordinates and thin-plate splines and then back to the organisms' picture for interpretation of these patterns as explanations. The reader who perseveres with the technic and the praxis of these preliminary chapters will be richly rewarded with the vistas that their joint application offers across the whole range of pattern studies of organismal form.

1.2 Our Basic Orientation: From Arithmetic to Understanding

Relying on this miscellany of strategies and tactics from so many different disciplines, inside biology or outside it, this book attempts to teach a particular kind of biological explanation, the kind arising from good pattern analyses of good numerical data about organismal form. By "good pattern analyses" I mean the highly structured numerical summaries (often quite clever or detailed) wherein crucial summary quantities match the features that we already know qualitatively to characterize the evolutionary history, life history, physiology, ecology, or pathology of the organisms we are measuring. If a study is about locomotion, energy consumption or transformation, growth, sex dimorphism, predation, reproduction, or selection, then the features of a good pattern analysis should align with those same conceptualizations. By "good

numerical data" I mean numbers generated by carefully calibrated machines, operating on physical principles, aggregated over samples whose systematic features correspond to the varieties of "natural kinds" with the aid of which we have some hope of "carving Nature at the joints," as Plato's Socrates famously put it (*Phaedrus* 265e). The word "good" here should be taken in the context of the craft knowledge possessed by the experienced biometrician or human biologist. Or perhaps that should be the phrase "good enough," as in "good enough to convince your colleagues" (for the equivalent in physics, see Krieger, 2012).

The purpose of the book is to sharpen your inferential skills insofar as they pertain to this general bioscientific task of bringing good pattern analyses to bear on good biological data toward our shared goal of heightened bioscientific understanding. Out of the signals detected by the pattern analyses, our aim is to build the explanations that lead to new *questions*, to promising *interventions* (on a human body, a farm, a forest, an ocean, or any other ecosystem), or to deeper academic or public *understanding* of how we became the seven-billion-odd fecund, sentient, frequently rational large mammals who, as I write this, have just outweighed the ants as the most successful eusocial organisms in the history of the planet, at least if success is measured by biomass (E. O. Wilson, 2012).

An especially good pattern analysis will not only meet these criteria but also model the multiple and protean appearances of *noise* in the data. (Here the word "noise" means, roughly, "every aspect of variability that you are uninterested in explaining, at least at this level of measurement.") It is in this way that noise enters into our reasoning here also. If you have taken a statistics class before, you are already acquainted with the commonest quantification of noise in statistical inference, which is the "standard error" of an average or a comparison of averages due to *random sampling*. But this theme, along with its accompanying liturgy of *statistical significance testing*, plays only a very minor role in the explanatory techniques and tools introduced here (see also Mosteller, 1968). We will be much more interested in the ways that measured variables of the same conceptual entity align or misalign, agree or disagree within the same animal or human specimen. And of course the appropriate model for multiple measures of the "same thing" is for them to agree, or, in another jargon, to correlate very highly, not to vary cluelessly with respect to one another. Except in a context of strong and previously confirmed theory, no null hypothesis is likely to play any important role in an organismal study. In fact, the converse is likelier to be true: if the null hypothesis is a serious alternative to the theory in which you are interested as an explanation of the data set you are facing, then you may be reading the wrong book. You should

instead be reading the technical manuals that might help you upgrade your instruments, or sharpening your sample descriptions until the appropriate pattern engine can find enough contrast between the signal and the noise to get a numerically reliable grip on the explanatory purpose of your scientific activity. Or maybe you are in the wrong discipline entirely, one that is not so pervasively characterized by systems-level organization and its variation as are the evolved, developed forms of the organisms we study.

This emphasis is different from that of other textbooks in our area, which, when they refer to regression or correlation, typically refer to purposes of "prediction" rather than those of explanation. For instance, in the opening sentence of his textbook on applied regression analysis (a technique that will be of central concern here), Weisberg, 2005) expresses the current conventional wisdom quite clearly without using the word "explanation" even once:

> Regression analysis answers questions about the dependence of a response variable on one or more predictors, including prediction of future values of a response, discovering which predictors are important, and estimating the impact of changing a predictor or a treatment on the value of the response. (p. xiii)

I don't agree – those are not the questions that regression answers, and they are not the biologist's most important questions anyway. In this book, regression analysis will not be argued to be capable of answering questions about "dependence" or "discovery," nor will its "predictions" be of any particular interest or importance in most of the scientific contexts we will be considering. Regression analysis, the way it will be taught here, is only arithmetic, mere numbers. On its own it cannot "predict future values," "discover which predictors are important," or "estimate the impact of changing a predictor or a treatment." Scientific meaning, if any is to obtain, must be based on axioms of causation and control, numerical stability, details of sample design, conditions of observation or intervention, choice(s) of measurements to be made, and, crucially, the route that runs from your data through all your arithmetic to arrive at a *persuasive numerical explanation,* the main scientific goal privileged in this book.

The background I am hoping you bring to the study of this book, then, is not the usual sketchy "introductory statistics" syllabus that focused on elementary probability theory and "significance testing" as part of the customary baccalaureate curriculum in many areas, but instead a background in biometrics or organismal biology that focused on the uses of quantification in the course of pursuing reliable scientific insights pertinent to studies at all levels (not just the molecular or genetic) of the complex systems we call organisms, including