Statistics for Social and Behavioral Sciences

Living Standards Analytics

Development through the Lens of Household Survey Data

Statistics for Social and Behavioral Sciences

For further volumes:
http://www.springer.com/series/3463

Dominique Haughton • Jonathan Haughton

Living Standards Analytics

Development through the Lens of Household Survey Data

Springer

Dominique Haughton
Department of Mathematical Sciences
Bentley College
Waltham, MA, USA
dhaughton@bentley.edu

Jonathan Haughton
Department of Economics
Suffolk University
Boston, MA, USA
jhaughto@beaconhill.org

ISBN 978-1-4614-0384-5 e-ISBN 978-1-4614-0385-2
DOI 10.1007/978-1-4614-0385-2
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011934800

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To our parents
Monique and Paul Boudier
and
Helen and Joe Haughton
and to our daughter Isabelle

Preface

The Gallup Organization polls a thousand people every day. The Thailand Statistical Office interviews 3,000 households, using detailed surveys, every month. The amount of digital information doubles every 18 months.

We are, to use a headline from *The Economist*, facing a data deluge. What a contrast to the time when Nobel prize winner Wassily Leontief (1971), in his Presidential Address to the American Economic Association, could complain about a plethora of theory and a dearth of data, and call for a shift to "large-scale factual analysis" (p. 5).

The earliest analysis of household survey data – going back at least to the pioneering work of Seebohm Rowntree (1901) – was largely confined to tabulations. Starting in 1980, the World Bank's Living Standards Measurement Survey project boosted the pace and quality of household survey data gathered in Less-Developed Countries; 89 of the surveys may be downloaded from its Web site, but hundreds more such surveys are now available. By 2002 the project had generated 135 technical papers. This second wave emphasized the use of graphical and regression techniques, nicely summed up in the essential volume by Angus Deaton, *The Analysis of Household Surveys: A Microeconometric Approach to Economic Development* (1987).

We are now experiencing a third wave, with the increasing application of an ever-broadening array of analytical tools – such as directed acyclic graphs (DAGs), Kohonen maps, and propensity score matching – in addition to refinements to regression.

The purpose of this book is to introduce, discuss, illustrate, and evaluate the colorful palette of analytical techniques that can be applied to the analysis of household survey data, with an emphasis on the innovations of the past decade or so. It is conceived as an antidote to an overly narrow view of what constitutes legitimate empirical work, and reflects our own preferences as methodological eclectics.

The term "analytics" means the science of analysis, and in the business world – from which we have borrowed the term – it denotes the use of data, often in large quantities, to improve decision making. We use the term in its widest sense, as the harnessing of data, particularly from household surveys, to improve policy

recommendations. It is a large canvas, ranging from the mainstream econometric approach of testing and subsequently revising the sharp lines of model-generated hypotheses – what Deaton (2010, p. 4) calls the hypothetico-deductive method – to the soft brush strokes of some of the atheoretical techniques of data mining and exploratory data analysis. Both painting styles have their place in the gallery of analytics.

This is a gateway book. Most of the chapters begin by introducing a methodological or policy problem, to motivate the subsequent discussion of relevant methods. They then summarize the relevant techniques, and draw on examples – many of them from our own work – and aim to convey a sense of the potential, but also the strengths and weaknesses, of those techniques. The idea is to provide enough detail to allow the reader to take the next steps, but not so much detail as to get bogged down.

To be exhaustive would be too exhausting. For example, we introduce Kohonen maps in Chap. 6, explain how they function, and work through an example. The interested reader will then be well positioned to dig deeper, into a field where more than 5,000 articles have been published.

In writing this book, we have three main audiences in mind. The first is graduate students in statistics, economics, policy analysis, and social sciences, especially, but certainly not exclusively, those interested in the challenges of economic development in the Third World. We would be delighted if this book opens the reader to a handful of new ideas: skim the book, alight on the pages that catch one's fancy, and return to it regularly as a reference and a fount of ideas.

Our second target group is academics, who will likely be very conversant with some of the material in the book, but would appreciate a quick *tour d'horizon* to familiarize them with other interesting, and potentially useful, techniques. This is a book, like Deaton's *Analysis of Household Surveys*, that can serve as a reference work, to be taken down from the shelf and perused from time to time.

Our third audience is practitioners, by whom we mean anyone who works closely with survey data, whether in statistics offices, think tanks, research units, international organizations, central banks, NGOs, businesses – the list is long. We know, from teaching online and internationally, that there are many who, having left the university environment, are not sure how to keep up with new technical developments; we believe the book will help, because it introduces the techniques and ideas without getting too lost in the technical detail.

The Substance

We begin the book with a consideration of graphical methods, because this is often the first step when we are trying to develop a feel for our data. Graphs can be revealing, and they can be helpful in presenting our findings. We start by discussing how to produce a useful histogram, and its continuous-valued cousin, the kernel density. Boxplots are also easy to use and especially helpful when we want to

compare the essential features of two or more distributions side by side. The chapter also includes some discussion of violin plots, scatterplots, and bag plots, before turning to presentational graphics. We agree with Gelman et al. (2002) that graphs could productively be used more often when presenting scientific results: The beautiful bubble plot in Fig. 1.13 contains more information than its apparent simplicity would suggest. The final section of Chap. 1 looks at maps, which can now be produced remarkably quickly and easily; the cartogram in Fig. 1.17 shows the distribution of child mortality worldwide, and instantly conveys the locus of the problem.

After graphics comes regression, which we survey in Chap. 2. Seasoned econometricians and other quantitative researchers can skip this chapter, but it is our experience that regression is sufficiently subtle, and the ideas sufficiently slippery, that one needs a quick review of the material on a regular basis. We note the main problems faced in regression, including measurement error, omitted variable bias, multicolinearity, heteroscedasticity, adjustments for clustered data, outliers, and simultaneity, and suggest ways in which these may be dealt with. Thus the chapter includes a discussion of, among other things, instrumental variables, and quantile regression. It is a whistle stop tour, which is exactly what most of us need.

Household survey data almost never come from simple random samples, and in Chap. 3 we address the issues related to sampling, first reviewing the main types – simple, stratified, cluster – and then presenting the essentials of how to determine an appropriate sample size while recognizing the need to trade off sampling with nonsampling errors. We show how to incorporate sample design into the computation of summary statistics – using Stata, the statistical package that we have used most over the years – and summarize the debate on whether to use weights in regression. The last two sections of the chapter ask how best to survey hard-to-reach groups, such as migrants – the main focus of a recent survey in the two main cities of Vietnam – and groups such as jazz players, or prostitutes, where respondent-driven sampling has been quite successful.

In Chap. 4 we move beyond linear regression, first by making the linear specification more flexible, and then by using nonparametric methods to fit curves. This segues into an explanation of multivariate adaptive regression spline (MARS) models, which we apply to a model of changes in consumption spending in Vietnam between 1993 and 1998. We also discuss classification and regression tree (CART) models; both CART and MARS are particularly good at exploring the data for nonlinearities and interactions. We have used a CART model with some success as a first step in helping us specify the functional form of a parametric model of the determinants of short-term malnutrition in Vietnam.

Much of our interest in working with living standards survey data arises from our desire to say something useful for policy purposes. This requires us to be able to say, "if you do X, then Y will happen," which is a causal statement. The question of causality, and more specifically how to conceive of and measure causal statements, is the subject of Chap. 5. The experimentalist school focuses on measuring the "effects of causes," where possible using randomized experiments to try to determine whether microcredit raises spending or flip charts improve exam performance.

The structuralist school worries that the outcomes of experiments leave us with an insufficient understanding of the underlying causal mechanisms, and urge us to pay attention to unearthing the "causes of effects," which may then be generalized to other situations and applied to policy. Taking its cue from Edward Tufte, who famously wrote that "correlation is not causation but it sure is a hint," the causal inference school, seeks to measure causality using a combination of correlations and logic. This approach is essentially mechanical, and the results are usually shown in the form of directed acyclic graphs (DAGs). This is unfamiliar terrain for most economists and policy analysts, which is why we devote much of the chapter to explaining how DAGs are constructed and what we might learn from them.

We often group data, for instance looking at income by gender, region, or quintile. In Chap. 6 we explore in more detail how observations may be clustered. This is an exploratory process, traditionally conducted with hierarchical or non-hierarchical clustering, which can produce beautiful graphs. It is also possible to incorporate more statistical structure using latent class models. The second half of the chapter introduces Kohonen maps, which have become very popular: They typically group observations on a two-dimensional grid, and present the results in the form of gorgeous "maps" – all of which we explain and illustrate here.

In approaching any scientific question, or looking at any data, we almost always have at least some idea of what we expect. If the data showed that richer households bought fewer cars, or poorer households eat more caviar, we would be shocked. Bayesian analysis provides a formal framework for incorporating these prior beliefs, in contrast to the more standard frequentist approach that either ignores them entirely, or locks them into rigid models. Chapter 7 provides an introduction to Bayesian analysis, setting out the ideas, the approach, and an example, and then addressing the problems of eliciting priors, applying posterior predictive checking, combining models in the form of Bayesian model averaging, and determining the appropriate sample size for a survey. This is not the easiest chapter in the book – the intrinsic difficulty of the subject helps explain its still-limited spread beyond trained statisticians – but it is likely to be one of the more useful for nonstatistician readers.

We are rediscovering geography, and recognizing once again that what happens in one area can influence what happens nearby. The presence of spatial dependence has implications for how to specify and estimate regression models – most commonly through the use of spatial weights matrices that measure the strength of the contiguity effects. We illustrate the use of these techniques in Chap. 8, drawing on a study of the spatial pattern of unemployment in the Midi-Pyrénées region of France, where we also present an algorithm for choosing among different types of spatial models.

Although it is still comparatively rare, increasing numbers of household surveys are based on panels, where households are surveyed repeatedly over time. In Chap. 9 we show how panel data can allow for more precise inference, and in many cases can help us tackle the knotty problem of unobserved heterogeneity: if households differ in ways we cannot observe, but these differences – in ability or drive, for instance – do not vary over time, then differenced data can sweep

away such effects, laying bare the relationships that we are usually interested in measuring. We illustrate this with an example in which we try to measure the effect on income of loans extended under the Thailand Village Fund, which burst onto the scene in 2002 and by 2004 had become the largest microcredit scheme in the world. Still, panel data are not a panacea; attrition bias can be a problem, and even without attrition, panels become less representative over time.

One of the most important uses of household survey data is to measure poverty, and vulnerability to poverty. Chapter 10 reviews this field, starting with the choice of a measure of well-being, through the construction of a poverty line, to the choice of a summary measure of poverty. We then discuss the robustness of poverty measures, focusing on sampling and measurement error, and explaining the notion of stochastic dominance. After a section in which we consider the problems peculiar to international comparisons of poverty, we consider ways in which vulnerability to poverty – defined as the probability that a household will be poor in the future – may be measured.

We return to an essentially technical issue in Chap. 11, where we look at bootstrapping. This is especially useful when we need to estimate the standard error of a measure – such as the Sen–Shorrocks–Thon index of poverty – and where an analytical formula is not available. The technique can be powerful, especially where the data come from complex samples, and is increasingly straightforward to implement; we illustrate this with an example in which we create a histogram of bootstrapped changes in the poverty rate in Vietnam between 1993 and 1998.

Does a program work? Was a project effective? These are questions addressed by impact evaluation, where we try to compare the actual outcomes, for those who have been "treated," with a counterfactual, which is our estimate of what would have happened in the absence of the program or project. The traditional gold standard is experimental design, or randomization, but in Chap. 12 we show that even this is not without its limitations. It is much more common to use quasi-experimental methods, of which the most popular are propensity score matching, double differences, and instrumental variables. For each of these we set out the principles, consider an example, and review both the strengths and weaknesses. This is a relatively long and detailed chapter, but it has proven to be effective when teaching impact evaluation to graduate students in economics.

Household survey data mainly come from large, complex questionnaires administered to relatively small samples of perhaps 5,000–10,000. This allows one to conduct the analysis at the level of a country or broad region, but not at the level of a small county or district. Yet we would often like to measure, for instance, poverty rates at a "small-area" level, the better to target spending to alleviate poverty. In Chap. 13 we discuss how to do this, first describing a basic synthetic regression model, and then explaining how one might estimate a two-level, or even multilevel, model with random effects. This chapter applies the methods to Vietnam, and includes two elegant maps that result from the analysis.

Perhaps it is fitting that the last chapter in the book, Chap. 14, looks at duration models. In many cases, the time dimension is central to the analysis, such as

the interval between one birth and the next, or the time spent unemployed. We introduce the Kaplan–Meier estimator, which allows for an exploratory analysis of duration data, and move on to the Cox proportional hazards model, parametric regression models, and mixture models of two Weibull regressions. As always, this chapter is designed to help the reader take the first steps – enough for the first draft of a solid research paper, even if lifting it to the level required for scholarly publication will always call for digging a bit deeper.

Where We Stand

We come to this book with different perspectives – one schooled in economics where the mindset is one of "model first, then test," the other more comfortable with data mining and letting the numbers speak "for themselves." The tension between these approaches runs throughout the book, and we see this as a virtue. One of us is skeptical that directed acyclic graphs are useful in helping us understand how the world really works, and thinks that the main virtue of Kohonen maps is that they are pretty. The other has yet to find an instrumental variable that looks compelling, and thinks that a lot of highfalutin theory is "nonsense on stilts." We do not try to resolve these debates – we are reminded of the observation by George Box that "essentially, all models are wrong, but some are useful" – but instead set out the techniques and ideas, to help the reader develop an informed opinion.

Together, we have over 50 years of experience working with household datasets, and have written over 200 papers, articles, and reports, over 80 of them in scholarly journals. This book is our take on what we find to be most useful, or at least intriguing or innovative; it also contains what we would like our students to know.

We are grateful to all of those who helped us on the way to this book. All of our more than 150 co-authors have at least some claim to intellectual parentage. Glenn Jenkins started the ball rolling in 1979 by interesting one of us in using survey data to address a practical development problem, in this case whether to build small-scale irrigation projects in Malaysia. In 1994, Mark Sidel encouraged us to work with the General Statistics Office in Hanoi; this, and the ongoing support from Nguyen Phong of the GSO, explains why so many of the examples in this book are drawn from the various living standards surveys undertaken in Vietnam.

We would like to thank our institutions – Bentley University and Suffolk University – for providing research support and sabbatical leaves that helped us get the book written. We are grateful to John Kimmel for trusting us with the project, and waiting patiently for it to progress, and to Marc Strauss for taking up the baton; to Dan Westbrook for reviewing an early draft; and to Maria Skaletsky, Sunida Susantud, Bayar Tumennasan, and Jason Wells for very helpful comments.

References

Deaton, Angus. 1997. *The analysis of household surveys: A microeconometric approach to development policy*. Baltimore: Johns Hopkins University Press.

The Economist. 2010. Special report on managing information. February 27.

Gelman, A., C. Pasarica, and R. Dodhia. 2002. Let's practice what we preach: Turning tables into graphs. *The American Statistician*, 56(2):121–130.

Leontief, Wassily. 1971. Theoretical assumptions and nonobserved facts. *American Economic Review*, 61(1):1–7.

Rowntree, Seebohm. 1901. *Poverty: A study of town life*. London: Macmillan.

About the Authors

Dominique Haughton (PhD, MIT 1983) is Professor of Mathematical Sciences at Bentley University in Waltham, Massachusetts, near Boston, and Affiliated Researcher at the Université Toulouse 1, France. Her major areas of interest are applied statistics, statistics and marketing, the analysis of living standards surveys, data mining, and model selection. She is the editor-in-chief of *Case Studies in Business, Industry and Government Statistics (CSBIGS)*, and has published over 50 articles in scholarly journals, including *The American Statistician, Annals of Statistics, Sankhya, Communications in Statistics,* and *Statistica Sinica*. In 2011, she was elected a Fellow of the American Statistical Association.

Jonathan Haughton (PhD, Harvard 1983) is Professor of Economics at Suffolk University, and Senior Economist at the Beacon Hill Institute for Public Policy, both in Boston. A specialist in the areas of economic development, international trade, and taxation, and a prize-winning teacher, he has lectured, taught, or conducted research in over a score of countries on five continents. His *Handbook on Poverty and Inequality* (with Shahidur Khandker) was published by the World Bank in 2009, his articles have appeared in over 30 scholarly journals, and he has written numerous book chapters and over a hundred reports.

Contents

Chapter 1
Graphical Methods

1.1 Introduction

It is tempting, but wrong, to believe that graphical techniques have little to offer for serious researchers in economics, statistics, or policy analysis. Their true power comes from the ability of the eye to discern patterns in a graph that are not clearly evident from lists of numbers or tabulated statistics. In Tufte's pithy phrase, "graphics reveal data" (Tufte 2001, p. 13).

We explore this theme in the chapter, beginning with the use of basic exploratory graphical methods in Sect. 1.2, considering presentational graphics in Sect. 1.3, and introducing some more recent techniques, including maps, in Sect. 1.4.

With data in hand, the most productive first step is often to explore the data graphically. These graphs do not have to be especially polished and beautiful; rather, they need to be easy to produce and thoroughly informative, a visual scratch pad where we use the power of graphics to get a sense of the shape of variables and the interactions among them.

Following Tufte (2001), the point can be emphasized elegantly with the help of Anscombe's quartet – four data sets, reproduced in Table 1.1, that may be summarized by the same linear model, and where the mean values of the X and Y variables are the same in each case. Yet a graphical display of the data sets (Fig. 1.1) demonstrates how very different they are. Real data do not usually yield such coherent or clear patterns, but a good initial graphical analysis can easily come up with surprises – showing outliers, suggesting a need to use a mixture of distributions, or raising questions about how variables are related.

Graphical techniques are also exceptionally useful in presenting the results of one's analysis, and we agree with Gelman et al. (2002) that they are typically underutilized for this purpose. But presentational graphics require an approach that is quite different from that of exploratory graphics: they serve to communicate ideas to others, and so they need to be more beautiful and more carefully constructed.

D. Haughton and J. Haughton, *Living Standards Analytics*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4614-0385-2_1,

Table 1.1 Anscombe's quartet of data sets

I		II		III		IV	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
9.0	**7.5**	**9.0**	**7.5**	**9.0**	**7.5**	**9.0**	**7.5**

Notes: Each data set has 11 observations; the means are shown in the bottom row. Every regression line is $Y = 3+0.5X$; the standard error of the slope coefficient is 0.118 and its t-statistic is 4.24. In every case, $R^2 = 0.67$

Source: Anscombe 1973

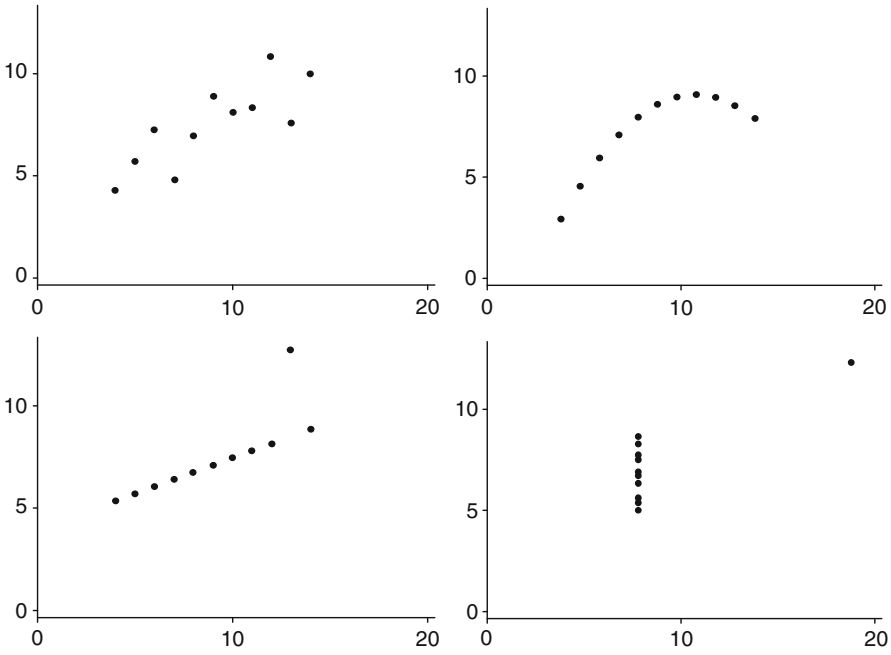

Fig. 1.1 Scatterplot of Anscombe's quartet (*Note*: Data from Table 1.1)

In Sect. 1.3 we review some of the key principles of graphical design, drawing heavily on the work of Tufte (2001), and suggest a few ways in which graphs could be used to make standard tabular presentations more effective.

1.2 Exploratory Graphical Methods

1.2.1 Histograms

A good place to start any analysis is with the most basic of visual techniques. Consider Fig. 1.2, which shows a simple frequency distribution (histogram) of birth weights of children born in Vietnam in 1992–1993. The data come from the Vietnam Living Standards Survey of 1992–1993, which surveyed 4,800 households nationwide and collected information on birth weights for 1,687 children. The graph represented the first step in an analysis by Sarah Bales (1999) of the determinants of low birth weights, and was generated using Stata.[1]

A baby is typically defined as being underweight if he or she weighs less than 2.5 kg at birth. Thus the histogram in Fig. 1.2 alerts us to a problem: an implausibly large number of births are heaped into the 2.5 kg category (and the 3.0, 3.5, and 4.0 kg categories). Indeed, 10.1% of the births were reported as weighing less than 2.5 kg and a further 10.7% as weighing exactly 2.5 kg! The rounding error matters here; the weight of some babies has presumably been rounded up to 2.5 kg, and in other cases the weight has been rounded down to 2.5 kg. So, while it is clear that more than 10.1% of babies are born underweight, but fewer than 20.8%, it is not clear whether it is preferable to define "underweight" as $w < 2.5$ or $w \leq 2.5$ (where w refers to the weight of the baby in kilos). The solution chosen by Bales (1999) was to use both definitions; fortunately, she found that the exact definition of underweight made relatively little difference to the direction and strength of the determinants of low birth weights.

Like a stethoscope, a histogram appears to be a simple tool, but it takes some practice to make it work effectively. The key choice that has to be made is that of the number of classes ("bins") into which to group the data or, alternatively, the width of each class, and this choice is as much a matter of art as of science.

A histogram aims to lay bare the distribution of the underlying data, and the classification of data into bins serves to filter out some of the noise. This is illustrated in Fig. 1.3, which displays four histograms showing the number of individuals covered by the 1998 Vietnam Living Standards Survey, broken down by age. The bottom right panel of Fig. 1.3 has just ten bins, and hints at a unimodal distribution dominated by the large proportion of individuals in the 10–20 age

[1] For a tutorial-based introduction to Stata, with examples that use easily accessible household survey data from Bangladesh, see Appendixes 1 and 2 of Haughton and Khandker (2009).

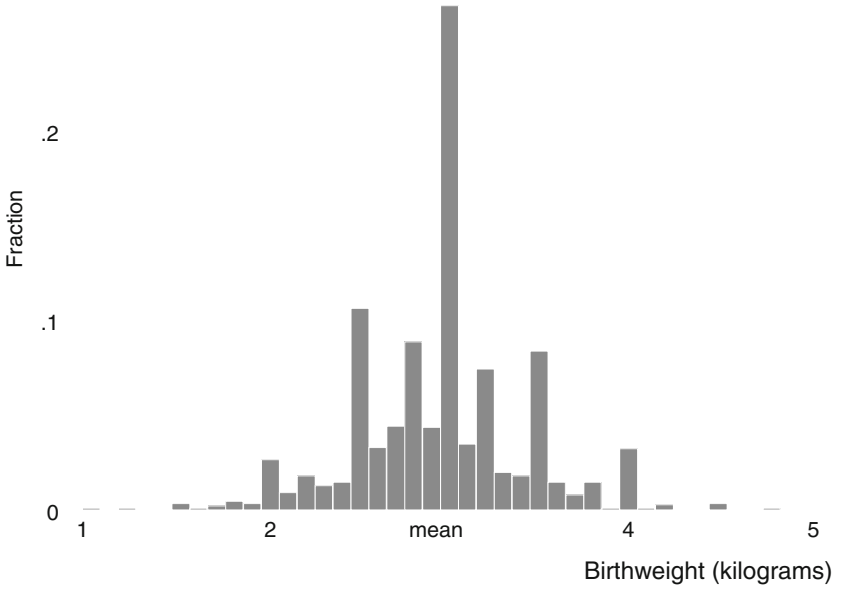

Fig. 1.2 Histogram of birth weights in Vietnam, 1993 (*Source*: Vietnam Living Standards Survey, 1993)

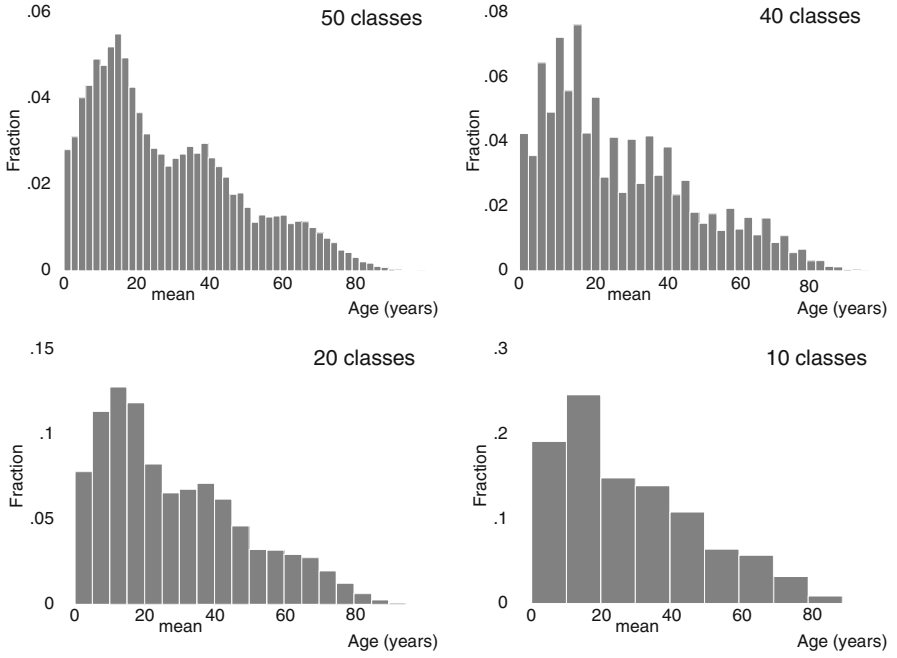

Fig. 1.3 Histograms of ages (in years) for individuals surveyed in the Vietnam Living Standards Survey of 1997–1998, with differing numbers of classes ("bins")

bracket. The top right panel uses 40 bins, and is clearly unsatisfactory; the age interval 0–100 is divided into 40 equal classes, so those aged 0, 1, or 2 are in the first bracket, those aged 3 or 4 in the next bracket, those aged 5, 6 or 7 in the third bin, and so on, in a saw-toothed fashion. This particular problem arises because the age data are integer rather than continuous, but even continuous data are often subject to heaping, as we saw in Fig. 1.2.

We get closer to a sensible pattern with a 20-bin histogram, which suggests a second peak. This is confirmed by the very informative histogram in the top left panel of Fig. 1.3, which uses 50 classes. Turned on its side, this gives half of a population "pyramid." Note, in this context, the dramatic reduction in the number of births since they peaked in about 1972 (i.e., 16 years prior to the 1998 survey), and the shortage of those aged roughly 45–60 – casualties of war, as they would have been in the armed forces in the years up to 1975 – and their children (Haughton 2000).

Are there better ways to choose bin widths other than trial and error? Freedman and Diaconis (1981) argue that if the histogram is to serve as a density estimator, then an appropriate rule for determining bin width is

$$\mathrm{BW}_{\mathrm{FD}}(x) = 2 \times \mathrm{IQR}(x) \times n^{-1/3}, \tag{1.1}$$

where $\mathrm{BW}_{\mathrm{FD}}(x)$ is the Freedman and Diaconis bin width for variable x, $\mathrm{IQR}(x)$ is the interquartile range of x, and n is the number of observations. With more observations we can afford to have narrower bins; with greater variation in x the bins need to be wider.

Other rules have been suggested. For instance, Wand (1997) proposes starting with a "zero-stage rule" that sets the bin width as

$$\mathrm{BW}_{\mathrm{W}}(x) = 3.49 \times \left[\min\left\{s, \frac{\mathrm{IQR}(x)}{1.349}\right\}\right] \times n^{-1/3}, \tag{1.2}$$

where s is the standard deviation of the sample.

In the example in Fig. 1.3, the Freedman–Diaconis rule gives a bin width of 1.9, implying 52 bins, while the Wand rule generates a bin width of 2.3, implying 43 bins. Neither rule gives results that are as clean as those with 50 bins, but they would filter the data nicely if one were using truly continuous (rather than integer) data.

Statistical software packages try to help the user by starting with sensible guesses of the appropriate number of bins. Microsoft Excel sets the number of bins equal to $\sqrt{n}$ (rounded to the next integer) or 50, whichever is the smallest. Stata uses an only slightly more complex default, which is

$$\text{Number of bins} = \min[50, \min\{\sqrt{n}, 10 \times \ln(n)/\ln(10)\}]. \tag{1.3}$$

This sometimes works well, but usually some further exploration is called for to produce a sensible histogram.

1.2.2 Kernel Densities

A histogram provides a discretized, nonparametric approximation to the underlying density, but it has three drawbacks: it is not smooth, it depends on the bin widths, and it is sensitive to the choice of end points of the bins. So it is often more useful, or at least more elegant, to work with a smoothed version. This is achieved by estimating a kernel density.

Suppose we have a dataset $X_1, X_2, \ldots, X_n$, and array the observations on the horizontal axis of a graph. We are interested in estimating the density, $f(x)$, at any given point x. A natural way to measure the density is by measuring the concentration of observed data points that are in the vicinity of x, say in the interval $x \pm h$, where h is the bandwidth half length. As we move x and its associated interval rightwards along the horizontal axis, we drop points to the left and pick up new observations on the right. The effect on the total number of observations is gradual, hence the smoothing effect.

The process we have described here may be formalized. It generates the naïve (or rectangular) estimate of the density at x, given by

$$\hat{f}_N(x) = \frac{1}{hn} \sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right) \equiv \frac{1}{hn} \sum_{i=1}^{n} W(z), \tag{1.4}$$

where

$$w(z) = \begin{cases} 1/2 & \text{if } |z|<1 \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

so that

$$\hat{f}_N(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} I(|z| \leq 1), \tag{1.6}$$

where $I(\cdot)$ is an indicator function that takes on a value of 1 if the bracketed expression is true, and zero otherwise.

More generally, we may define a kernel density estimator as

$$\hat{f}(x) = \frac{1}{hn} \left[\sum_{i=1}^{n} K(z) \right], \tag{1.7}$$

where $K(z)$ is the kernel function, calibrated so that the estimator integrates to 1. The naïve estimator puts an equal weight on all the observations in the interval $x \pm h$ when estimating the density of x, which is why the kernel function in this case is referred to as "rectangular." However, it is usually more satisfactory to use a symmetric function that puts more weight on values of X_i that are closer to x, and progressively less weight on values further from x. The widely used Epanechnikov

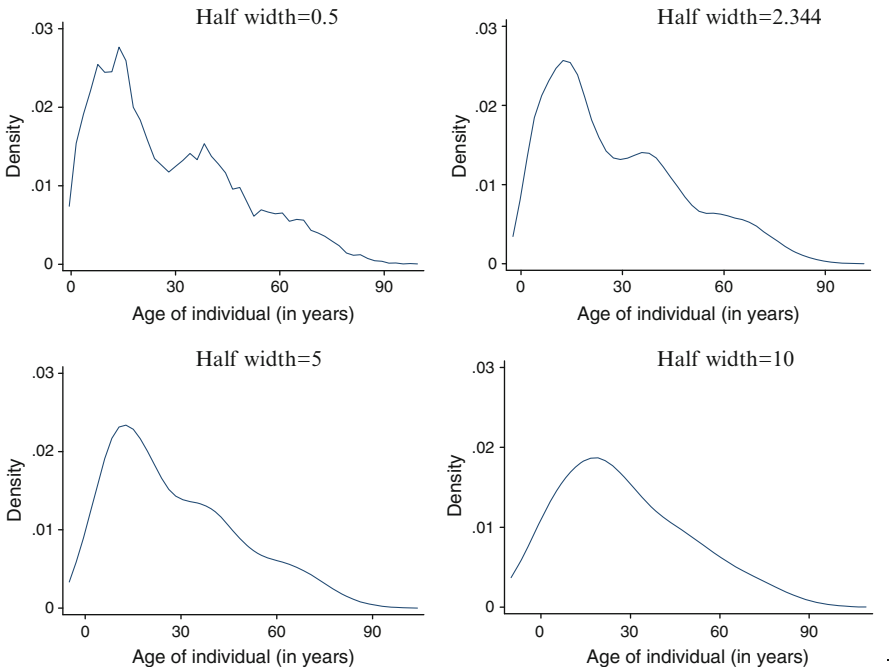

Fig. 1.4 Kernel densities of ages (in years) for individuals surveyed in the Vietnam Living Standards Survey of 1997–1998, with differing half-widths, including the Stata default value of 2.344

kernel is a concave quadratic function with maximum weight at x and zero weights at $x \pm h$. Formally,

$$K_{\mathrm{E}}(z) = \frac{3}{4} \times \left[1 - \frac{1}{5}z^2\right] \Big/ \sqrt{5} \times \mathrm{I}(|z| < \sqrt{5}). \tag{1.8}$$

The principal virtue of the Epanechnikov kernel is that it is the most efficient in minimizing the mean integrated squared error, which is the difference between the true and estimated densities (Stata 2010; Silverman 1986). The Gaussian kernel is also popular, and is given by

$$K_{\mathrm{G}}(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}. \tag{1.9}$$

In fitting a kernel density we have to choose both the kernel function itself and the bandwidth h. It is generally agreed that the important decision concerns the choice of h, just as the choice of bin width is central to the construction of a good histogram. If h is too wide, the kernel density filters the data too much, and potentially valuable information is lost – compare the bottom right panel of Fig. 1.4, which has a wide bandwidth, with the top right panel, where the bandwidth