1

Introduction

The explicit recognition of uncertainty is central to the statistical sciences. Notions such as prior information, probability models, likelihood, standard errors and confidence limits are all intended to formalize uncertainty and thereby make allowance for it. In simple situations, the uncertainty of an estimate may be gauged by analytical calculation based on an assumed probability model for the available data. But in more complicated problems this approach can be tedious and difficult, and its results are potentially misleading if inappropriate assumptions or simplifications have been made.

For illustration, consider Table 1.1, which is taken from a larger tabulation (Table 7.4) of the numbers of AIDS reports in England and Wales from mid-1983 to the end of 1992. Reports are cross-classified by diagnosis period and length of reporting delay, in three-month intervals. A blank in the table corresponds to an unknown (as yet unreported) entry. The problem was to predict the states of the epidemic in 1991 and 1992, which depend heavily on the values missing at the bottom right of the table.

The data support the assumption that the reporting delay does not depend on the diagnosis period. In this case a simple model is that the number of reports in row j and column k of the table has a Poisson distribution with mean $\mu_{jk} = \exp(\alpha_j + \beta_k)$. If all the cells of the table are regarded as independent, then the total number of unreported diagnoses in period j has a Poisson distribution with mean

$$\sum_k \mu_{jk} = \exp(\alpha_j) \sum_k \exp(\beta_k),$$

where the sum is over columns with blanks in row j. The eventual total of as yet unreported diagnoses from period j can be estimated by replacing α_j and β_k by estimates derived from the incomplete table, and thence we obtain the predicted total for period j. Such predictions are shown by the solid line in

Diagnosis period		Reporting delay interval (quarters):									Total reports to end
Year	Quarter	0†	1	2	3	4	5	6	···	≥14	of 1992
1988	1	31	80	16	9	3	2	8	···	6	174
	2	26	99	27	9	8	11	3	···	3	211
	3	31	95	35	13	18	4	6	···	3	224
	4	36	77	20	26	11	3	8	···	2	205
1989	1	32	92	32	10	12	19	12	···	2	224
	2	15	92	14	27	22	21	12	···	1	219
	3	34	104	29	31	18	8	6	···		253
	4	38	101	34	18	9	15	6	···		233
1990	1	31	124	47	24	11	15	8	···		281
	2	32	132	36	10	9	7	6	···		245
	3	49	107	51	17	15	8	9	···		260
	4	44	153	41	16	11	6	5	···		285
1991	1	41	137	29	33	7	11	6	···		271
	2	56	124	39	14	12	7	10			263
	3	53	175	35	17	13	11				306
	4	63	135	24	23	12					258
1992	1	71	161	48	25						310
	2	95	178	39							318
	3	76	181								273
	4	67									133

Table 1.1 Numbers of AIDS reports in England and Wales to the end of 1992 (De Angelis and Gilks, 1994) extracted from Table 7.4. A † indicates a reporting delay less than one month.

Figure 1.1, together with the observed total reports to the end of 1992. How good are these predictions?

It would be tedious but possible to put pen to paper and estimate the prediction uncertainty through calculations based on the Poisson model. But in fact the data are much more variable than that model would suggest, and by failing to take this into account we would believe that the predictions are more accurate than they really are. Furthermore, a better approach would be to use a semiparametric model to smooth out the evident variability of the increase in diagnoses from quarter to quarter; the corresponding prediction is the dotted line in Figure 1.1. Analytical calculations for this model would be very unpleasant, and a more flexible line of attack is needed. While more than one approach is possible, the one that we shall develop based on computer simulation is both flexible and straightforward.

Purpose of the Book

Our central goal is to describe how the computer can be harnessed to obtain reliable standard errors, confidence intervals, and other measures of uncertainty for a wide range of problems. The key idea is to resample from the original data — either directly or via a fitted model — to create replicate datasets, from

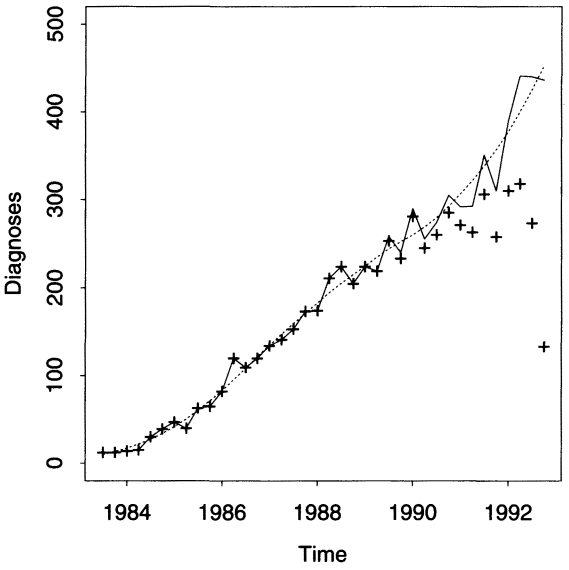

Figure 1.1 Predicted quarterly diagnoses from a parametric model (solid) and a semiparametric model (dots) fitted to the AIDS data, together with the actual totals to the end of 1992 (+).

which the variability of the quantities of interest can be assessed without long-winded and error-prone analytical calculation. Because this approach involves repeating the original data analysis procedure with many replicate sets of data, these are sometimes called *computer-intensive methods*. Another name for them is *bootstrap methods*, because to use the data to generate more data seems analogous to a trick used by the fictional Baron Munchausen, who when he found himself at the bottom of a lake got out by pulling himself up by his bootstraps. In the simplest nonparametric problems we do literally sample from the data, and a common initial reaction is that this is a fraud. In fact it is not. It turns out that a wide range of statistical problems can be tackled this way, liberating the investigator from the need to oversimplify complex problems. The approach can also be applied in simple problems, to check the adequacy of standard measures of uncertainty, to relax assumptions, and to give quick approximate solutions. An example of this is random sampling to estimate the permutation distribution of a nonparametric test statistic.

It is of course true that in many applications we can be fairly confident in a particular parametric model and the standard analysis based on that model. Even so, it can still be helpful to see what can be inferred without particular parametric model assumptions. This is in the spirit of *robustness of validity* of the statistical analysis performed. Nonparametric bootstrap analysis allows us to do this.

3	5	7	18	43	85	91	98	100	130	230	487

Table 1.2 Service hours between failures of the air-conditioning equipment in a Boeing 720 jet aircraft (Proschan, 1963).

Despite its scope and usefulness, resampling must be carefully applied. Unless certain basic ideas are understood, it is all too easy to produce a solution to the wrong problem, or a bad solution to the right one. Bootstrap methods are intended to help avoid tedious calculations based on questionable assumptions, and this they do. But they cannot replace clear critical thought about the problem, appropriate design of the investigation and data analysis, and incisive presentation of conclusions.

In this book we describe how resampling methods can be used, and evaluate their performance, in a wide range of contexts. Our focus is on the methods and their practical application rather than on the underlying theory, accounts of which are available elsewhere. This book is intended to be useful to the many investigators who want to know how and when the methods can safely be applied, and how to tell when things have gone wrong. The mathematical level of the book reflects this: we have aimed for a clear account of the key ideas without an overload of technical detail.

Examples

Bootstrap methods can be applied both when there is a well-defined probability model for data and when there is not. In our initial development of the methods we shall make frequent use of two simple examples, one of each type, to illustrate the main points.

Example 1.1 (Air-conditioning data) Table 1.2 gives $n = 12$ times between failures of air-conditioning equipment, for which we wish to estimate the underlying mean or its reciprocal, the failure rate. A simple model for this problem is that the times are sampled from an exponential distribution.

The dotted line in the left panel of Figure 1.2 is the cumulative distribution function (CDF)

$$F_\mu(y) = \begin{cases} 0, & y \le 0, \\ 1 - \exp(-y/\mu), & y > 0, \end{cases}$$

for the fitted exponential distribution with mean μ set equal to the sample average, $\bar{y} = 108.083$. The solid line on the same plot is the nonparametric equivalent, the empirical distribution function (EDF) for the data, which places equal probabilities $n^{-1} = 0.08\dot{3}$ at each sample value. Comparison of the two curves suggests that the exponential model fits reasonably well. An alternative view of this is shown in the right panel of the figure, which is an exponential

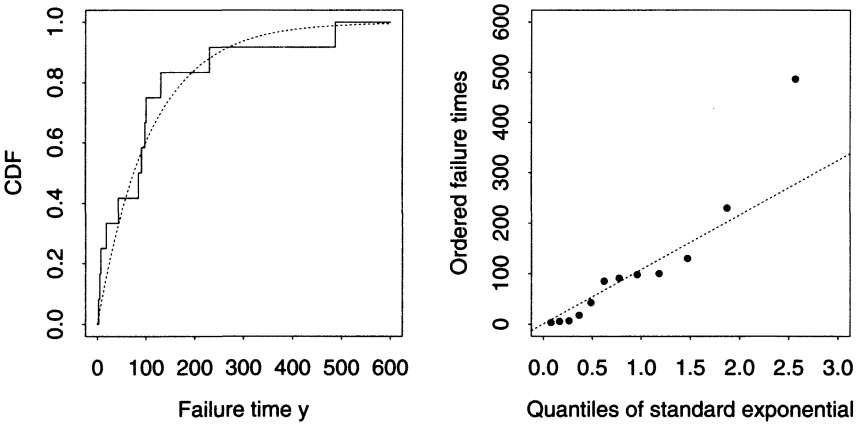

Figure 1.2 Summary displays for the air-conditioning data. The left panel shows the EDF for the data, $\hat{F}$ (solid), and the CDF of a fitted exponential distribution (dots). The right panel shows a plot of the ordered failure times against exponential quantiles, with the fitted exponential model shown as the dotted line.

Q-Q plot — a plot of ordered data values $y_{(j)}$ against the standard exponential quantiles

$$F_{\mu}^{-1}\left(\frac{j}{n+1}\right)\Bigg|_{\mu=1} = -\log\left(1-\frac{j}{n+1}\right).$$

Although these plots suggest reasonable agreement with the exponential model, the sample is rather too small to have much confidence in this. In the data source the more general gamma model with mean μ and index κ is used; its density is

$$f_{\mu,\kappa}(y) = \frac{1}{\Gamma(\kappa)}\left(\frac{\kappa}{\mu}\right)^{\kappa} y^{\kappa-1}\exp(-\kappa y/\mu), \qquad y>0, \quad \mu,\kappa>0. \tag{1.1}$$

For our sample the estimated index is $\hat{\kappa} = 0.71$, which does not differ significantly ($P = 0.29$) from the value $\kappa = 1$ that corresponds to the exponential model. Our reason for mentioning this will become apparent in Chapter 2.

Basic properties of the estimator $T = \bar{Y}$ for μ are easy to obtain theoretically under the exponential model. For example, it is easy to show that T is unbiased and has variance μ^2/n. Approximate confidence intervals for μ can be calculated using these properties in conjunction with a normal approximation for the distribution of T, although this does not work very well: we can tell this because $\bar{Y}/\mu$ has an exact gamma distribution, which leads to exact confidence limits. Things are more complicated under the more general gamma model, because the index κ is only estimated, and so in a traditional approach we would use approximations — such as a normal approximation for the distribution of T, or a chi-squared approximation for the log likelihood ratio statistic.

The parametric simulation methods of Section 2.2 can be used alongside these approximations, to diagnose problems with them, or to replace them entirely. ■

Example 1.2 (City population data) Table 1.3 reports $n = 49$ data pairs, each corresponding to a city in the United States of America, the pair being the 1920 and 1930 populations of the city, which we denote by u and x. The data are plotted in Figure 1.3. Interest here is in the ratio of means, because this would enable us to estimate the total population of the USA in 1930 from the 1920 figure. If the cities form a random sample with (U, X) denoting the pair of population values for a randomly selected city, then the total 1930 population is the product of the total 1920 population and the ratio of expectations $\theta = \mathrm{E}(X)/\mathrm{E}(U)$. This ratio is the parameter of interest.

In this case there is no obvious parametric model for the joint distribution of (U, X), so it is natural to estimate θ by its empirical analog, $T = \bar{X}/\bar{U}$, the ratio of sample averages. We are then concerned with the uncertainty in T. If we had a plausible parametric model — for example, that the pair (U, X) has a bivariate lognormal distribution — then theoretical calculations like those in Example 1.1 would lead to bias and variance estimates for use in a normal approximation, which in turn would provide approximate confidence intervals for θ. Without such a model we must use nonparametric analysis. It is still possible to estimate the bias and variance of T, as we shall see, and this makes normal approximation still feasible, as well as more complex approaches to setting confidence intervals. ■

Example 1.1 is special in that an exact distribution is available for the statistic of interest and can be used to calculate confidence limits, at least under the exponential model. But for parametric models in general this will not be true. In Section 2.2 we shall show how to use parametric simulation to obtain approximate distributions, either by approximating moments for use in normal approximations, or — when these are inaccurate — directly.

In Example 1.2 we make no assumptions about the form of the data disribution. But still, as we shall show in Section 2.3, simulation can be used to obtain properties of T, even to approximate its distribution. Much of Chapter 2 is devoted to this.

Layout of the Book

Chapter 2 describes the properties of resampling methods for use with single samples from parametric and nonparametric models, discusses practical matters such as the numbers of replicate datasets required, and outlines delta methods for variance approximation based on different forms of jackknife. It

Table 1.3 Populations in thousands of $n = 49$ large US cities in 1920 (u) and in 1930 (x) (Cochran, 1977, p. 152).

u	x	u	x	u	x
138	143	76	80	67	67
93	104	381	464	120	115
61	69	387	459	172	183
179	260	78	106	66	86
48	75	60	57	46	65
37	63	507	634	121	113
29	50	50	64	44	58
23	48	77	89	64	63
30	111	64	77	56	142
2	50	40	60	40	64
38	52	136	139	116	130
46	53	243	291	87	105
71	79	256	288	43	61
25	57	94	85	43	50
298	317	36	46	161	232
74	93	45	53	36	54
50	58				

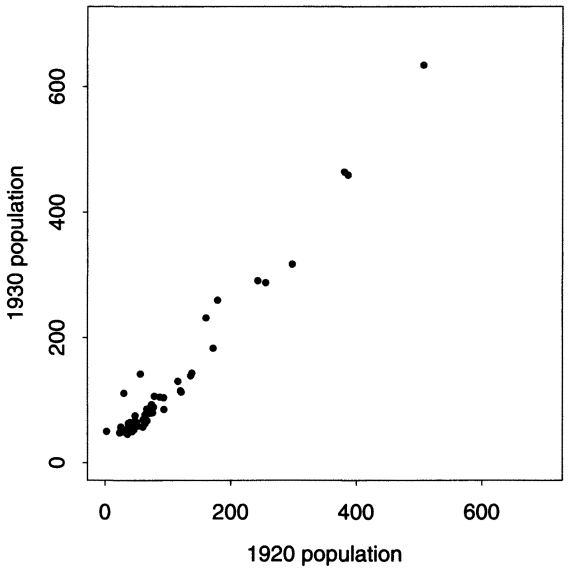

Figure 1.3 Populations of 49 large United States cities (in 1000s) in 1920 and 1930.

also contains a basic discussion of confidence intervals and of the ideas that underlie bootstrap methods.

Chapter 3 outlines how the basic ideas are extended to several samples, semiparametric and smooth models, simple cases where data have hierarchical structure or are sampled from a finite population, and to situations where data are incomplete because censored or missing. It goes on to discuss how the simulation output itself may be used to detect problems — so-called bootstrap diagnostics — and how it may be useful to bootstrap the bootstrap.

In Chapter 4 we review the basic principles of significance testing, and then describe Monte Carlo tests, including those using Markov Chain simulation, and parametric bootstrap tests. This is followed by discussion of nonparametric permutation tests, and the more general methods of semi- and nonparametric bootstrap tests. A double bootstrap method is detailed for improved approximation of P-values.

Confidence intervals are the subject of Chapter 5. After outlining basic ideas, we describe how to construct simple confidence intervals based on simulations, and then go on to more complex methods, such as the studentized bootstrap, percentile methods, the double bootstrap and test inversion. The main methods are compared empirically in Section 5.7, then there are brief accounts of confidence regions for multivariate parameters, and of prediction intervals.

The three subsequent chapters deal with more complex problems. Chapter 6 describes how the basic resampling methods may be applied in linear regression problems, including tests for coefficients, prediction analysis, and variable selection. Chapter 7 deals with more complex regression situations: generalized linear models, other nonlinear models, semi- and nonparametric regression, survival analysis, and classification error. Chapter 8 details methods appropriate for time series, spatial data, and point processes.

Chapter 9 describes how variance reduction techniques such as balanced simulation, control variates, and importance sampling can be adapted to yield improved simulations, with the aim of reducing the amount of simulation needed for an answer of given accuracy. It also shows how saddlepoint methods can sometimes be used to avoid simulation entirely.

Chapter 10 describes various semiparametric versions of the likelihood function, the ideas underlying which are closely related to resampling methods. It also briefly outlines a Bayesian version of the bootstrap.

Chapters 2–10 contain problems intended to reinforce the reader's understanding of both methods and theory, and in some cases problems develop topics that could not be included in the text. Some of these demand a knowledge of moments and cumulants, basic facts about which are sketched in the Appendix.

The book also contains practicals that apply resampling routines written in

the S language to sets of data. The practicals are intended to reinforce the ideas in each chapter, to supplement the more theoretical problems, and to give examples on which readers can base analyses of their own data.

It would be possible to give different sorts of course based on this book. One would be a "theoretical" course based on the problems and another an "applied" course based on the practicals; we prefer to blend the two.

Although a library of routines for use with the statistical package `SPlus` is bundled with it, most of the book can be read without reference to particular software packages. Apart from the practicals, the exception to this is Chapter 11, which is a short introduction to the main resampling routines, arranged roughly in the order with which the corresponding ideas appear in earlier chapters. Readers intending to use the bundled routines will find it useful to work through the relevant sections of Chapter 11 before attempting the practicals.

Notation

Although we believe that our notation is largely standard, there are not enough letters in the English and Greek alphabets for us to be entirely consistent. Greek letters such as θ, β and ν generally denote parameters or other unknowns, while α is used for error rates in connection with significance tests and confidence sets. English letters X, Y, Z, and so forth are used for random variables, which take values x, y, z. Thus the estimator T has observed value t, which may be an estimate of the unknown parameter θ. The letter V is used for a variance estimate, and the letter p for a probability, except for regression models, where p is the number of covariates. Script letters such as $\mathcal{N}$ are used to denote sets.

Probability, expectation, variance and covariance are denoted $\Pr(\cdot)$, $\mathrm{E}(\cdot)$, $\mathrm{var}(\cdot)$ and $\mathrm{cov}(\cdot,\cdot)$, while the joint cumulant of Y_1, Y_1Y_2 and Y_3 is denoted $\mathrm{cum}(Y_1, Y_1Y_2, Y_3)$. We use $I\{A\}$ to denote the indicator random variable, which takes values one if the event A is true and zero otherwise. A related function is the Heaviside function

$$H(u) = \begin{cases} 0, & u < 0, \\ 1, & u \geq 0. \end{cases}$$

We use $\#\{A\}$ to denote the number of elements in the set A, and $\#\{A_r\}$ for the number of events A_r that occur in a sequence $A_1, A_2, \ldots$. We use $\doteq$ to mean "is approximately equal to", usually corresponding to asymptotic equivalence as sample sizes tend to infinity, $\sim$ to mean "is distributed as" or "is distributed according to", $\dot{\sim}$ to mean "is distributed approximately as", $\overset{iid}{\sim}$ to mean "is a sample of independent identically distributed random variables from", while $\equiv$ has its usual meaning of "is equivalent to".

The data values in a sample of size n are typically denoted by $y_1, \ldots, y_n$, the observed values of the random variables $Y_1, \ldots, Y_n$; their average is $\bar{y} = n^{-1} \sum y_j$.

We mostly reserve Z for random variables that are standard normal, at least approximately, and use Q for random variables with other (approximately) known distributions. As usual $N(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2, while z_α is often the α quantile of the standard normal distribution, whose cumulative distribution function is $\Phi(\cdot)$.

The letter R is reserved for the number of replicate simulations. Simulated copies of a statistic T are denoted T_r^*, $r = 1, \ldots, R$, whose ordered values are $T_{(1)}^* \le \cdots \le T_{(R)}^*$. Expectation, variance and probability calculated with respect to the simulation distribution are written $\Pr^*(\cdot)$, $\mathrm{E}^*(\cdot)$ and $\mathrm{var}^*(\cdot)$.

Where possible we avoid boldface type, and rely on the context to make it plain when we are dealing with vectors or matrices; a^T denotes the matrix transpose of a vector or matrix a.

We use PDF, CDF, and EDF as shorthand for "probability density function", "cumulative distribution function", and "empirical distribution function". The letters F and G are used for CDFs, and f and g are generally used for the corresponding PDFs. An exception to this is that f_{rj}^* denotes the frequency with which y_j appears in the rth resample.

We use MLE as shorthand for "maximum likelihood estimate" or sometimes "maximum likelihood estimation".

The end of each example is marked ■, and the end of each algorithm is marked •.