Part I

Classic Statistical Inference

1

Algorithms and Inference

Statistics is the science of learning from experience, particularly experience that arrives a little bit at a time: the successes and failures of a new experimental drug, the uncertain measurements of an asteroid's path toward Earth. It may seem surprising that any one theory can cover such an amorphous target as "learning from experience." In fact, there are *two* main statistical theories, Bayesianism and frequentism, whose connections and disagreements animate many of the succeeding chapters.

First, however, we want to discuss a less philosophical, more operational division of labor that applies to both theories: between the *algorithmic* and *inferential* aspects of statistical analysis. The distinction begins with the most basic, and most popular, statistical method, averaging. Suppose we have observed numbers $x_1, x_2, \ldots, x_n$ applying to some phenomenon of interest, perhaps the automobile accident rates in the $n = 50$ states. The *mean*

$$\bar{x} = \sum_{i=1}^{n} x_i / n \tag{1.1}$$

summarizes the results in a single number.

How accurate is that number? The textbook answer is given in terms of the *standard error*,

$$\widehat{\text{se}} = \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n(n-1)) \right]^{1/2}. \tag{1.2}$$

Here *averaging* (1.1) is the algorithm, while the standard error provides an inference of the algorithm's accuracy. It is a surprising, and crucial, aspect of statistical theory that the same data that supplies an estimate can also assess its accuracy.[1]

[1] "Inference" concerns more than accuracy: speaking broadly, algorithms say what the statistician does while inference says why he or she does it.

Of course, $\widehat{\text{se}}$ (1.2) is itself an algorithm, which could be (and is) subject to further inferential analysis concerning *its* accuracy. The point is that the algorithm comes first and the inference follows at a second level of statistical consideration. In practice this means that algorithmic invention is a more free-wheeling and adventurous enterprise, with inference playing catch-up as it strives to assess the accuracy, good or bad, of some hot new algorithmic methodology.

If the inference/algorithm race is a tortoise-and-hare affair, then modern electronic computation has bred a bionic hare. There are two effects at work here: computer-based technology allows scientists to collect enormous data sets, orders of magnitude larger than those that classic statistical theory was designed to deal with; huge data demands new methodology, and the demand is being met by a burst of innovative computer-based statistical algorithms. When one reads of "big data" in the news, it is usually these algorithms playing the starring roles.

Our book's title, *Computer Age Statistical Inference*, emphasizes the tortoise's side of the story. The past few decades have been a golden age of statistical methodology. It hasn't been, quite, a golden age for statistical inference, but it has not been a dark age either. The efflorescence of ambitious new algorithms has forced an evolution (though not a revolution) in inference, the theories by which statisticians choose among competing methods. The book traces the interplay between methodology and inference as it has developed since the 1950s, the beginning of our discipline's computer age. As a preview, we end this chapter with two examples illustrating the transition from classic to computer-age practice.

1.1 A Regression Example

Figure 1.1 concerns a study of kidney function. Data points (x_i, y_i) have been observed for $n = 157$ healthy volunteers, with x_i the ith volunteer's `age` in years, and y_i a composite measure "`tot`" of overall function. Kidney function generally declines with `age`, as evident in the downward scatter of the points. The rate of decline is an important question in kidney transplantation: in the past, potential donors past `age` 60 were prohibited, though, given a shortage of donors, this is no longer enforced.

The solid line in Figure 1.1 is a *linear regression*

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1.3}$$

fit to the data by *least squares*, that is by minimizing the sum of squared

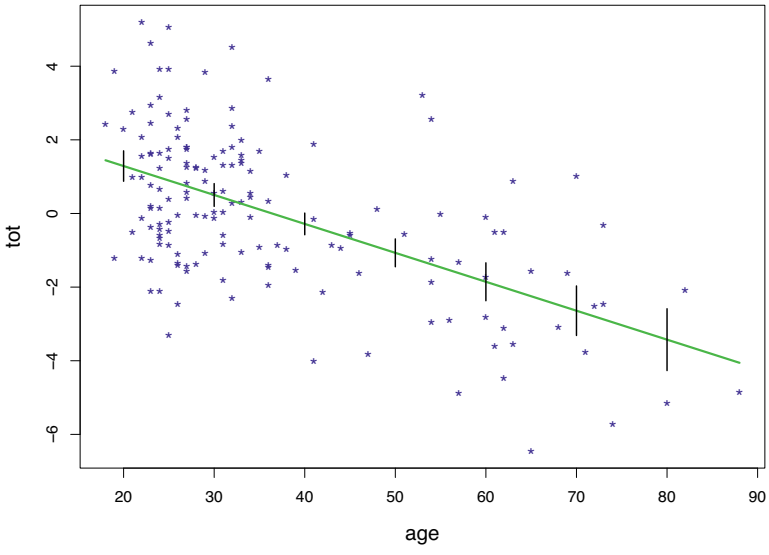

Figure 1.1 Kidney fitness `tot` vs `age` for 157 volunteers. The line is a linear regression fit, showing ±2 standard errors at selected values of `age`.

deviations

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{1.4}$$

over all choices of (β_0, β_1). The least squares algorithm, which dates back to Gauss and Legendre in the early 1800s, gives $\hat{\beta}_0 = 2.86$ and $\hat{\beta}_1 = -0.079$ as the least squares estimates. We can read off of the fitted line an estimated value of kidney fitness for any chosen `age`. The top line of Table 1.1 shows estimate 1.29 at `age` 20, down to −3.43 at `age` 80.

How accurate are these estimates? This is where inference comes in: an extended version of formula (1.2), also going back to the 1800s, provides the standard errors, shown in line 2 of the table. The vertical bars in Figure 1.1 are ± two standard errors, giving them about 95% chance of containing the true expected value of `tot` at each `age`.

That 95% coverage depends on the validity of the linear regression model (1.3). We might instead try a quadratic regression $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$, or a cubic, etc., all of this being well within the reach of pre-computer statistical theory.

Table 1.1 *Regression analysis of the kidney data; (1) linear regression estimates; (2) their standard errors; (3)* `lowess` *estimates; (4) their bootstrap standard errors.*

`age`	20	30	40	50	60	70	80
1. linear regression	1.29	.50	−.28	−1.07	−1.86	−2.64	−3.43
2. std error	.21	.15	.15	.19	.26	.34	.42
3. lowess	1.66	.65	−.59	−1.27	−1.91	−2.68	−3.50
4. bootstrap std error	.71	.23	.31	.32	.37	.47	.70

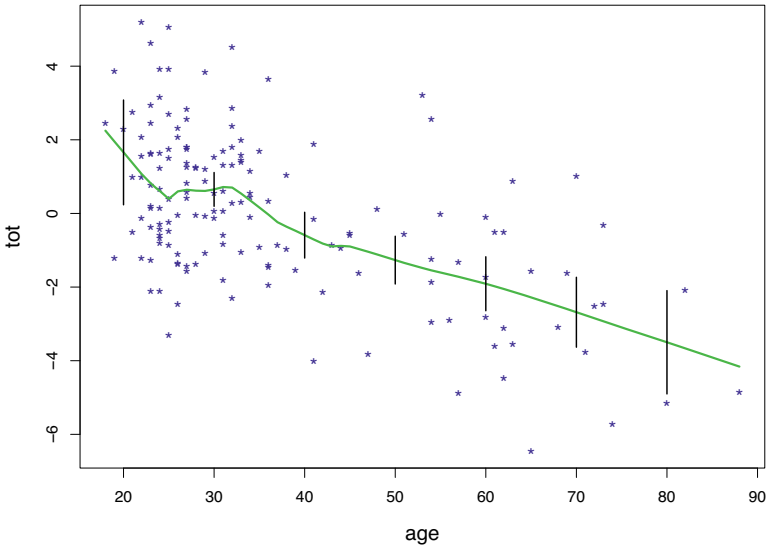

Figure 1.2 Local polynomial `lowess(x,y,1/3)` fit to the kidney-fitness data, with ±2 bootstrap standard deviations.

†1 A modern computer-based algorithm `lowess` produced the somewhat bumpy regression curve in Figure 1.2. The `lowess`† [2] algorithm moves its attention along the x-axis, fitting local polynomial curves of differing degrees to nearby (x, y) points. (The 1/3 in the call[3] `lowess(x,y,1/3)`

[2] Here and throughout the book, the numbered † sign indicates a technical note or reference element which is elaborated on at the end of the chapter.

[3] Here and in all our examples we are employing the language `R`, itself one of the key developments in computer-based statistical methodology.

determines the definition of local.) Repeated passes over the x-axis refine the fit, reducing the effects of occasional anomalous points. The fitted curve in Figure 1.2 is nearly linear at the right, but more complicated at the left where points are more densely packed. It is flat between ages 25 and 35, a potentially important difference from the uniform decline portrayed in Figure 1.1.

There is no formula such as (1.2) to infer the accuracy of the `lowess` curve. Instead, a computer-intensive inferential engine, the *bootstrap*, was used to calculate the error bars in Figure 1.2. A bootstrap data set is produced by resampling 157 pairs (x_i, y_i) from the original 157 *with replacement*, so perhaps (x_1, y_1) might show up twice in the bootstrap sample, (x_2, y_2) might be missing, (x_3, y_3) present once, etc. Applying `lowess` to the bootstrap sample generates a bootstrap replication of the original calculation.

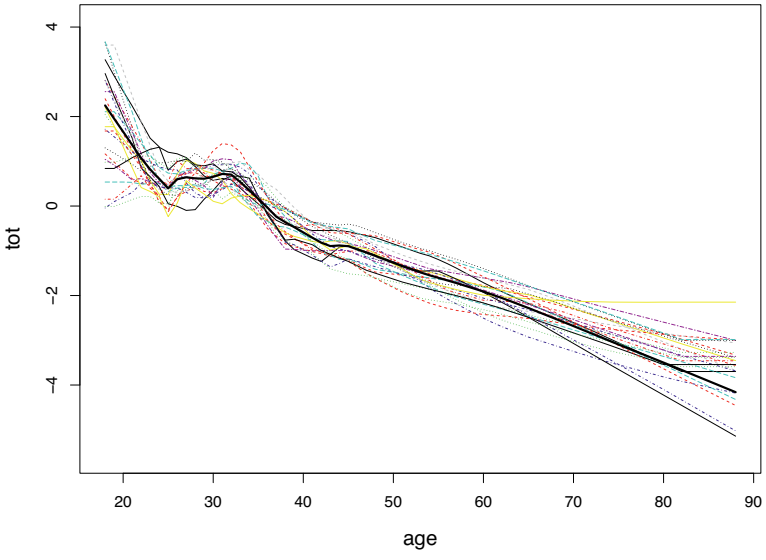

Figure 1.3 25 bootstrap replications of `lowess(x,y,1/3)`.

Figure 1.3 shows the first 25 (of 250) bootstrap `lowess` replications bouncing around the original curve from Figure 1.2. The variability of the replications at any one `age`, the *bootstrap standard deviation*, determined the original curve's accuracy. How and why the bootstrap works is discussed in Chapter 10. It has the great virtue of assessing estimation accu-

racy for *any* algorithm, no matter how complicated. The price is a hundred- or thousand-fold increase in computation, unthinkable in 1930, but routine now.

The bottom two lines of Table 1.1 show the `lowess` estimates and their standard errors. We have paid a price for the increased flexibility of `lowess`, its standard errors roughly doubling those for linear regression.

1.2 Hypothesis Testing

Our second example concerns the march of methodology and inference for *hypothesis testing* rather than estimation: 72 leukemia patients, 47 with `ALL` (acute lymphoblastic leukemia) and 25 with `AML` (acute myeloid leukemia, a worse prognosis) have each had genetic activity measured for a panel of 7,128 genes. The histograms in Figure 1.4 compare the genetic activities in the two groups for gene 136.

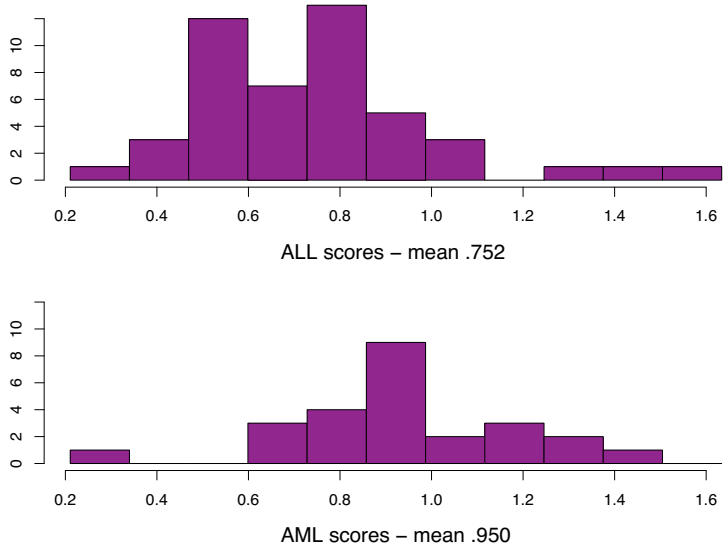

Figure 1.4 Scores for gene 136, leukemia data. Top `ALL` ($n = 47$), bottom `AML` ($n = 25$). A two-sample t-statistic $= 3.01$ with p-value $= .0036$.

The `AML` group appears to show greater activity, the mean values being

$$\overline{\texttt{ALL}} = 0.752 \quad \text{and} \quad \overline{\texttt{AML}} = 0.950. \tag{1.5}$$

Is the perceived difference genuine, or perhaps, as people like to say, "a statistical fluke"? The classic answer to this question is via a *two-sample t-statistic*,

$$t = \frac{\overline{\texttt{AML}} - \overline{\texttt{ALL}}}{\widehat{\text{sd}}}, \tag{1.6}$$

where $\widehat{\text{sd}}$ is an estimate of the numerator's standard deviation.[4]

Dividing by $\widehat{\text{sd}}$ allows us (under Gaussian assumptions discussed in Chapter 5) to compare the observed value of t with a standard "null" distribution, in this case a Student's t distribution with 70 degrees of freedom. We obtain $t = 3.01$ from (1.6), which would classically be considered very strong evidence that the apparent difference (1.5) is genuine; in standard terminology, "with two-sided significance level 0.0036."

A small significance level (or "p-value") is a statement of statistical surprise: something very unusual has happened if in fact there is no difference in gene 136 expression levels between **ALL** and **AML** patients. We are less surprised by $t = 3.01$ if gene 136 is just one candidate out of thousands that might have produced "interesting" results.

That is the case here. Figure 1.5 shows the histogram of the two-sample t-statistics for the panel of 7128 genes. Now $t = 3.01$ looks less unusual; 400 other genes have t exceeding 3.01, about 5.6% of them.

This doesn't mean that gene 136 is "significant at the 0.056 level." There are two powerful complicating factors:

1 Large numbers of candidates, 7128 here, will produce some large t-values even if there is really no difference in genetic expression between **ALL** and **AML** patients.
2 The histogram implies that in this study there is something wrong with the theoretical null distribution ("Student's t with 70 degrees of freedom"), the smooth curve in Figure 1.5. It is much too narrow at the center, where presumably most of the genes are reporting non-significant results.

We will see in Chapter 15 that a low *false-discovery rate*, i.e., a low chance of crying wolf over an innocuous gene, requires t exceeding 6.16 in the **ALL**/**AML** study. Only 47 of the 7128 genes make the cut. False-discovery-rate theory is an impressive advance in statistical inference, incorporating Bayesian, frequentist, and empirical Bayesian (Chapter 6) el-

[4] Formally, a standard error is the standard deviation of a summary statistic, and $\widehat{\text{sd}}$ might better be called $\widehat{\text{se}}$, but we will follow the distinction less than punctiliously here.

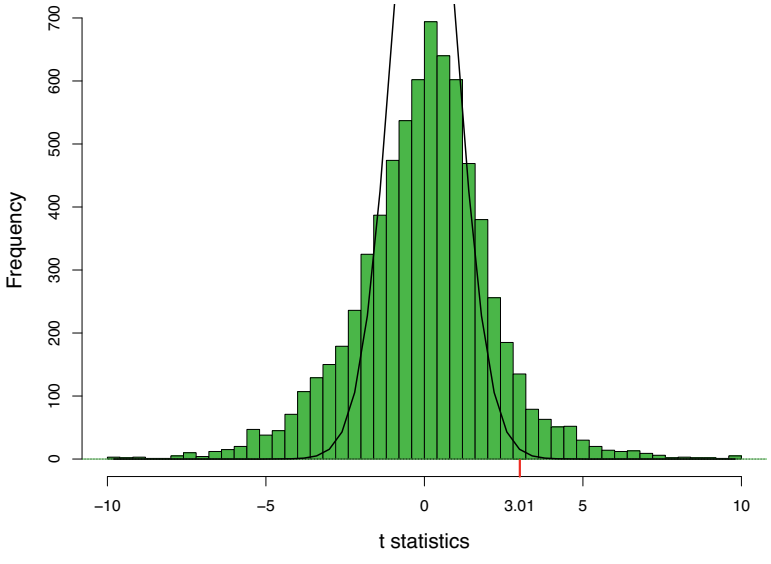

Figure 1.5 Two-sample t-statistics for 7128 genes, leukemia data. The smooth curve is the theoretical null density for the t-statistic.

ements. It was a *necessary* advance in a scientific world where computer-based technology routinely presents thousands of comparisons to be evaluated at once.

There is one more thing to say about the algorithm/inference statistical cycle. Important new algorithms often arise outside the world of professional statisticians: neural nets, support vector machines, and boosting are three famous examples. None of this is surprising. New sources of data, satellite imagery for example, or medical microarrays, inspire novel methodology from the observing scientists. The early literature tends toward the enthusiastic, with claims of enormous applicability and power.

In the second phase, statisticians try to locate the new metholodogy within the framework of statistical theory. In other words, they carry out the statistical inference part of the cycle, placing the new methodology within the known Bayesian and frequentist limits of performance. (Boosting offers a nice example, Chapter 17.) This is a healthy chain of events, good both for the hybrid vigor of the statistics profession and for the further progress of algorithmic technology.

1.3 Notes

Legendre published the least squares algorithm in 1805, causing Gauss to state that he had been using the method in astronomical orbit-fitting since 1795. Given Gauss' astonishing production of major mathematical advances, this says something about the importance attached to the least squares idea. Chapter 8 includes its usual algebraic formulation, as well as Gauss' formula for the standard errors, line 2 of Table 1.1.

Our division between algorithms and inference brings to mind Tukey's exploratory/confirmatory system. However the current algorithmic world is often bolder in its claims than the word "exploratory" implies, while to our minds "inference" conveys something richer than mere confirmation.

†1 [p. 6] `lowess` was devised by William Cleveland (Cleveland, 1981) and is available in the R statistical computing language. It is applied to the kidney data in Efron (2004). The kidney data originated in the nephrology laboratory of Dr. Brian Myers, Stanford University, and is available from this book's web site.

1.4 Exercises

1 1 Fit a cubic regression, as a function of age, to the `kidney` data of Figures 1.1 and 1.2, calculating estimates and standard errors at ages 20, 30, 40, 50, 60, 70, 80.
  2 How do the results compare with those in Table 1.1?

2 The `lowess` curve in Figure 1.2 has a flat spot between ages 25 and 35. Discuss how one might use bootstrap replications like those in Figure 1.3 to suggest whether the flat spot is genuine or just a statistical artifact.

3 Suppose that there were no differences between AML and ALL patients for any gene, so that t in (1.6) exactly followed a Student's t distribution with 70 degrees of freedom in all 7128 cases. *About* how big might you expect the largest observed t-value to be? *Hint:* $1/7128 = 0.00014$.

4 1 Perform 1000 nonparametric bootstrap replications of $\overline{\text{ALL}}$ (1.5). You can use program `bcanon` from the CRAN library "bootstrap" or type in Algorithm 10.1 on page 186.
  2 Do the same for $\overline{\text{AML}}$.
  3 Plot histograms of the results, and suggest an inference.

5 Statistical methods by their nature combine information from multiple persons or situations. A statement such as "the new treatment cured 82% of the cases in a study of 500 patients" attempts to learn the treatment's efficacy from its observed performance. Suppose you were thinking of

using this treatment. What might be arguments for and against taking "82%" seriously?