10

The Jackknife and the Bootstrap

A central element of frequentist inference is the *standard error*. An algorithm has produced an estimate of a parameter of interest, for instance the mean $\bar{x} = 0.752$ for the 47 **ALL** scores in the top panel of Figure 1.4. How accurate is the estimate? In this case, formula (1.2) for the standard deviation[1] of a sample mean gives estimated standard error

$$\widehat{\text{se}} = 0.040, \tag{10.1}$$

so one can't take the third digit of $\bar{x} = 0.752$ very seriously, and even the 5 is dubious.

Direct standard error formulas like (1.2) exist for various forms of averaging, such as linear regression (7.34), and for hardly anything else. Taylor series approximations ("device 2" of Section 2.1) extend the formulas to smooth functions of averages, as in (8.30). Before computers, applied statisticians needed to be Taylor series experts in laboriously pursuing the accuracy of even moderately complicated statistics.

The jackknife (1957) was a first step toward a computation-based, nonformulaic approach to standard errors. The bootstrap (1979) went further toward automating a wide variety of inferential calculations, including standard errors. Besides sparing statisticians the exhaustion of tedious routine calculations the jackknife and bootstrap opened the door for more complicated estimation algorithms, which could be pursued with the assurance that their accuracy would be easily assessed. This chapter focuses on standard errors, with more adventurous bootstrap ideas deferred to Chapter 11. We end with a brief discussion of accuracy estimation for robust statistics.

[1] We will use the terms "standard error" and "standard deviation" interchangeably.

10.1 The Jackknife Estimate of Standard Error

The basic applications of the jackknife apply to *one-sample problems*, where the statistician has observed an independent and identically distributed (iid) sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ from an unknown probability distribution F on some space $\mathcal{X}$,

$$x_i \overset{\text{iid}}{\sim} F \qquad \text{for } i = 1, 2, \ldots, n. \tag{10.2}$$

$\mathcal{X}$ can be anything: the real line, the plane, a function space.[2] A *real-valued* statistic $\hat{\theta}$ has been computed by applying some algorithm $s(\cdot)$ to $\boldsymbol{x}$,

$$\hat{\theta} = s(\boldsymbol{x}), \tag{10.3}$$

and we wish to assign a standard error to $\hat{\theta}$. That is, we wish to estimate the standard deviation of $\hat{\theta} = s(\boldsymbol{x})$ under sampling model (10.2).

Let $\boldsymbol{x}_{(i)}$ be the sample with x_i removed,

$$\boldsymbol{x}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)', \tag{10.4}$$

and denote the corresponding value of the statistic of interest as

$$\hat{\theta}_{(i)} = s(\boldsymbol{x}_{(i)}). \tag{10.5}$$

Then the *jackknife estimate of standard error* for $\hat{\theta}$ is

$$\widehat{\text{se}}_{\text{jack}} = \left[\frac{n-1}{n}\sum_1^n \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2\right]^{1/2}, \quad \text{with } \hat{\theta}_{(\cdot)} = \sum_1^n \hat{\theta}_{(i)}/n. \tag{10.6}$$

In the case where $\hat{\theta}$ is the mean $\bar{x}$ of real values $x_1, x_2, \ldots, x_n$ (i.e., $\mathcal{X}$ is an interval of the real line), $\hat{\theta}_{(i)}$ is their average excluding x_i, which can be expressed as

$$\hat{\theta}_{(i)} = (n\bar{x} - x_i)/(n-1). \tag{10.7}$$

Equation (10.7) gives $\hat{\theta}_{(\cdot)} = \bar{x}$, $\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} = (\bar{x} - x_i)/(n-1)$, and

$$\widehat{\text{se}}_{\text{jack}} = \left[\sum_{i=1}^n (x_i - \bar{x})^2/\left(n(n-1)\right)\right]^{1/2}, \tag{10.8}$$

exactly the same as the classic formula (1.2). This is no coincidence. The fudge factor $(n-1)/n$ in definition (10.6) was inserted to make $\widehat{\text{se}}_{\text{jack}}$ agree with (1.2) when $\hat{\theta}$ is $\bar{x}$.

[2] If $\mathcal{X}$ is an interval of the real line we might take F to be the usual cumulative distribution function, but here we will just think of F as any full description of the probability distribution for an x_i on $\mathcal{X}$.

The advantage of $\widehat{\text{se}}_{\text{jack}}$ is that definition (10.6) can be applied in an automatic way to *any* statistic $\hat{\theta} = s(\boldsymbol{x})$. All that is needed is an algorithm that computes $s(\cdot)$ for the deleted data sets $\boldsymbol{x}_{(i)}$. Computer power is being substituted for theoretical Taylor series calculations. Later we will see that the underlying inferential ideas—plug-in estimation of frequentist standard errors—haven't changed, only their implementation.

As an example, consider the kidney function data set of Section 1.1. Here the data consists of $n = 157$ points (x_i, y_i), with x = age and y = tot in Figure 1.1. (So the generic x_i in (10.2) now represents the pair (x_i, y_i), and F describes a distribution in the plane.) Suppose we are interested in the correlation between age and tot, estimated by the usual sample correlation $\hat{\theta} = s(\boldsymbol{x})$,

$$s(\boldsymbol{x}) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \Bigg/ \left[\sum_{1}^{n}(x_i - \bar{x})^2 \sum_{1}^{n}(y_i - \bar{y})^2\right]^{1/2}, \quad (10.9)$$

computed to be $\hat{\theta} = -0.572$ for the kidney data.

Applying (10.6) gave $\widehat{\text{se}}_{\text{jack}} = 0.058$ for the accuracy of $\hat{\theta}$. Nonparametric bootstrap computations, Section 10.2, also gave estimated standard error 0.058. The classic Taylor series formula looks quite formidable in this case,

$$\widehat{\text{se}}_{\text{taylor}} = \left\{\frac{\hat{\theta}^2}{4n}\left[\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} - \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{20}} - \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}}\right]\right\}^{1/2} \quad (10.10)$$

where

$$\hat{\mu}_{hk} = \sum_{i=1}^{n}(x_i - \bar{x})^h(y_i - \bar{y})^k/n. \quad (10.11)$$

It gave $\widehat{\text{se}} = 0.057$.

It is worth emphasizing some features of the jackknife formula (10.6).

- It is nonparametric; no special form of the underlying distribution F need be assumed.
- It is completely automatic: a single master algorithm can be written that inputs the data set $\boldsymbol{x}$ and the function $s(\boldsymbol{x})$, and outputs $\widehat{\text{se}}_{\text{jack}}$.
- The algorithm works with data sets of size $n-1$, not n. There is a hidden assumption of smooth behavior across sample sizes. This can be worrisome for statistics like the sample median that have a different definition for odd and even sample size.

- The jackknife standard error is upwardly biased as an estimate of the true standard error.[†] †1
- The connection of the jackknife formula (10.6) with Taylor series methods is closer than it appears. We can write

$$\widehat{\text{se}}_{\text{jack}} = \left[\frac{\sum_1^n D_i^2}{n^2}\right]^{1/2}, \qquad \text{where } D_i = \frac{\hat\theta_{(i)} - \hat\theta_{(\cdot)}}{1/\sqrt{n(n-1)}}. \tag{10.12}$$

As discussed in Section 10.3, the D_i are approximate *directional derivatives*, measures of how fast the statistic $s(\boldsymbol{x})$ is changing as we decrease the weight on data point x_i. So $\text{se}^2_{\text{jack}}$ is proportional to the sum of squared derivatives of $s(\boldsymbol{x})$ in the n component directions. Taylor series expressions such as (10.10) amount to doing the derivatives by formula rather than numerically.

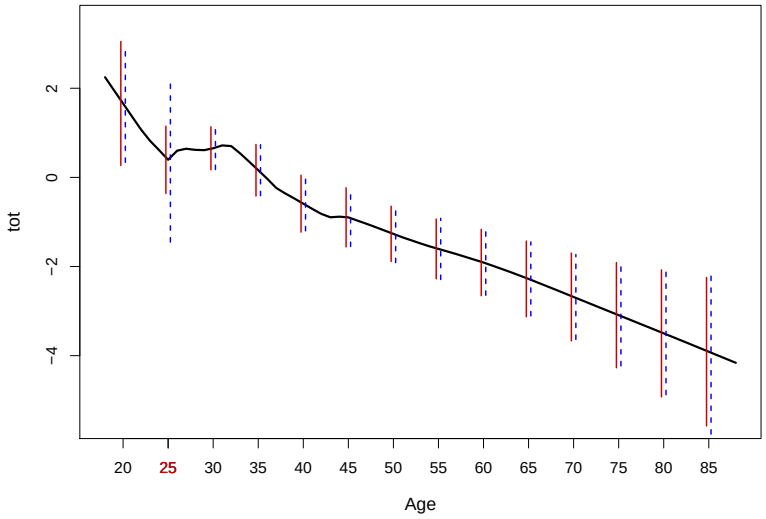

Figure 10.1 The `lowess` curve for the kidney data of Figure 1.2. Vertical bars indicate ±2 standard errors: *jackknife* (10.6) blue dashed; *bootstrap* (10.16) red solid. The jackknife greatly overestimates variability at age 25.

The principal weakness of the jackknife is its dependence on local derivatives. Unsmooth statistics $s(\boldsymbol{x})$, such as the kidney data `lowess` curve in Figure 1.2, can result in erratic behavior for $\widehat{\text{se}}_{\text{jack}}$. Figure 10.1 illustrates the point. The dashed blue vertical bars indicate ±2 jackknife standard er-

rors for the `lowess` curve evaluated at ages $20, 25, \ldots, 85$. For the most part these agree with the dependable bootstrap standard errors, solid red bars, described in Section 10.2. But things go awry at age 25, where the local derivatives greatly overstate the sensitivity of the `lowess` curve to global changes in the sample $\boldsymbol{x}$.

10.2 The Nonparametric Bootstrap

From the point of view of the bootstrap, the jackknife was a halfway house between classical methodology and a full-throated use of electronic computation. (The term "computer-intensive statistics" was coined to describe the bootstrap.) The frequentist standard error of an estimate $\hat{\theta} = s(\boldsymbol{x})$ is, ideally, the standard deviation we would observe by repeatedly sampling new versions of $\boldsymbol{x}$ from F. This is impossible since F is unknown. Instead, the bootstrap ("ingenious device" number 4 in Section 2.1) substitutes an estimate $\hat{F}$ for F and then estimates the frequentist standard by direct simulation, a feasible tactic only since the advent of electronic computation.

The bootstrap estimate of standard error for a statistic $\hat{\theta} = s(\boldsymbol{x})$ computed from a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ (10.2) begins with the notion of a *bootstrap sample*

$$\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*), \tag{10.13}$$

where each x_i^* is drawn randomly with equal probability and with replacement from $\{x_1, x_2, \ldots, x_n\}$. Each bootstrap sample provides a *bootstrap replication* of the statistic of interest,[3]

$$\hat{\theta}^* = s(\boldsymbol{x}^*). \tag{10.14}$$

Some large number B of bootstrap samples are independently drawn ($B = 500$ in Figure 10.1). The corresponding bootstrap replications are calculated, say

$$\hat{\theta}^{*b} = s(\boldsymbol{x}^{*b}) \qquad \text{for } b = 1, 2, \ldots, B. \tag{10.15}$$

The resulting bootstrap estimate of standard error for $\hat{\theta}$ is the empirical standard deviation of the $\hat{\theta}^{*b}$ values,

$$\widehat{\text{se}}_{\text{boot}} = \left[\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \hat{\theta}^{*\cdot}\right)^2 \Big/ (B-1)\right]^{1/2}, \quad \text{with } \hat{\theta}^{*\cdot} = \sum_{b=1}^{B} \hat{\theta}^{*b}/B. \tag{10.16}$$

[3] The star notation $\boldsymbol{x}^*$ is intended to avoid confusion with the original data $\boldsymbol{x}$, which stays fixed in bootstrap computations, and likewise $\hat{\theta}^*$ vis-a-vis $\hat{\theta}$.

Motivation for $\widehat{\text{se}}_{\text{boot}}$ begins by noting that $\hat{\theta}$ is obtained in two steps: first $\boldsymbol{x}$ is generated by iid sampling from probability distribution F, and then $\hat{\theta}$ is calculated from $\boldsymbol{x}$ according to algorithm $s(\cdot)$,

$$F \xrightarrow{\text{iid}} \boldsymbol{x} \xrightarrow{s} \hat{\theta}. \tag{10.17}$$

We don't know F, but we can estimate it by the *empirical probability distribution* $\hat{F}$ that puts probability $1/n$ on each point x_i (e.g., weight $1/157$ on each point (x_i, y_i) in Figure 1.2). Notice that a bootstrap sample $\boldsymbol{x}^*$ (10.13) *is an iid sample drawn from* $\hat{F}$, since then each $\boldsymbol{x}^*$ independently has equal probability of being any member of $\{x_1, x_2, \ldots, x_n\}$. It can be shown that $\hat{F}$ maximizes the probability of obtaining the observed sample $\boldsymbol{x}$ under all possible choices of F in (10.2), i.e., it is the *nonparametric MLE* of F.

Bootstrap replications $\hat{\theta}^*$ are obtained by a process analogous to (10.17),

$$\hat{F} \xrightarrow{\text{iid}} \boldsymbol{x}^* \xrightarrow{s} \hat{\theta}^*. \tag{10.18}$$

In the real world (10.17) we only get to see the single value $\hat{\theta}$, but the bootstrap world (10.18) is more generous: we can generate as many bootstrap replications $\hat{\theta}^{*b}$ as we want, or have time for, and directly estimate their variability as in (10.16). The fact that $\hat{F}$ approaches F as n grows large suggests, correctly in most cases, that $\widehat{\text{se}}_{\text{boot}}$ approaches the true standard error of $\hat{\theta}$.

The true standard deviation of $\hat{\theta}$, i.e., its standard error, can be thought of as a function of the probability distribution F that generates the data, say $\text{Sd}(F)$. Hypothetically, $\text{Sd}(F)$ inputs F and outputs the standard deviation of $\hat{\theta}$, which we can imagine being evaluated by independently running (10.17) some enormous number of times N, and then computing the empirical standard deviation of the resulting $\hat{\theta}$ values,

$$\text{Sd}(F) = \left[\sum_{j=1}^{N}\left(\hat{\theta}^{(j)} - \hat{\theta}^{(\cdot)}\right)^2 \Big/ (N-1)\right]^{1/2}, \quad \text{with } \hat{\theta}^{(\cdot)} = \sum_{1}^{N} \hat{\theta}^{(j)}/N. \tag{10.19}$$

The bootstrap standard error of $\hat{\theta}$ is the plug-in estimate

$$\widehat{\text{se}}_{\text{boot}} = \text{Sd}(\hat{F}). \tag{10.20}$$

More exactly, $\text{Sd}(\hat{F})$ is the *ideal bootstrap estimate* of standard error, what we would get by letting the number of bootstrap replications B go to infinity. In practice we have to stop at some finite value of B, as discussed in what follows.

As with the jackknife, there are several important points worth emphasizing about $\widehat{\text{se}}_{\text{boot}}$.

- It is completely automatic. Once again, a master algorithm can be written that inputs the data $\boldsymbol{x}$ and the function $s(\cdot)$, and outputs $\widehat{\text{se}}_{\text{boot}}$.
- We have described the *one-sample nonparametric bootstrap*. Parametric and multisample versions will be taken up later.
- Bootstrapping "shakes" the original data more violently than jackknifing, producing nonlocal deviations of $\boldsymbol{x}^*$ from $\boldsymbol{x}$. The bootstrap is more dependable than the jackknife for unsmooth statistics since it doesn't depend on local derivatives.
- $B = 200$ is usually sufficient† for evaluating $\widehat{\text{se}}_{\text{boot}}$. Larger values, 1000 or 2000, will be required for the bootstrap confidence intervals of Chapter 11. †2
- There is nothing special about standard errors. We could just as well use the bootstrap replications to estimate the expected absolute error $E\{|\hat{\theta} - \theta|\}$, or any other accuracy measure.
- Fisher's MLE formula (4.27) is applied in practice via

$$\widehat{\text{se}}_{\text{fisher}} = (n\mathcal{I}_{\hat{\theta}})^{-1/2}, \tag{10.21}$$

  that is, by plugging in $\hat{\theta}$ for θ after a theoretical calculation of se. The bootstrap operates in the same way at (10.20), though the plugging in is done before rather than after the calculation. The connection with Fisherian theory is more obvious for the parametric bootstrap of Section 10.4.

The jackknife is a completely frequentist device, both in its assumptions and in its applications (standard errors and biases). The bootstrap is also basically frequentist, but with a touch of the Fisherian as in the relation with (10.21). Its versatility has led to applications in a variety of estimation and prediction problems, with even some Bayesian connections.† †3
Unusual applications can also pop up for the jackknife; see the jackknife-after-bootstrap comment in the chapter endnotes.† †4

From a classical point of view, the bootstrap is an incredible computational spendthrift. Classical statistics was fashioned to minimize the hard labor of mechanical computation. The bootstrap seems to go out of its way to multiply it, by factors of $B = 200$ or 2000 or more. It is nice to report that all this computational largesse can have surprising data analytic payoffs.

The 22 students of Table 3.1 actually each took five tests, **mechanics**, **vectors**, **algebra**, **analytics**, and **statistics**. Table 10.1 shows

Table 10.1 *Correlation matrix for the student score data. The eigenvalues are 3.463, 0.660, 0.447, 0.234, and 0.197. The eigenratio statistic* $\hat{\theta} = 0.693$*, and its bootstrap standard error estimate is 0.075* ($B = 2000$).

	mechanics	vectors	algebra	analytics	statistics
mechanics	1.00	.50	.76	.65	.54
vectors	.50	1.00	.59	.51	.38
algebra	.76	.59	1.00	.76	.67
analysis	.65	.51	.76	1.00	.74
statistics	.54	.38	.67	.74	1.00

the sample correlation matrix and also its eigenvalues. The "eigenratio" statistic,

$$\hat{\theta} = \text{ largest eigenvalue/sum eigenvalues,} \tag{10.22}$$

measures how closely the five scores can be predicted by a single linear combination, essentially an IQ score for each student: $\hat{\theta} = 0.693$ here, indicating strong predictive power for the IQ score. How accurate is 0.693?

$B = 2000$ bootstrap replications (10.15) yielded bootstrap standard error estimate (10.16) $\widehat{\text{se}}_{\text{boot}} = 0.075$. (This was 10 times more bootstraps than necessary for $\widehat{\text{se}}_{\text{boot}}$, but will be needed for Chapter 11's bootstrap confidence interval calculations.) The jackknife (10.6) gave a bigger estimate, $\widehat{\text{se}}_{\text{jack}} = 0.083$.

Standard errors are usually used to suggest approximate confidence intervals, often $\hat{\theta} \pm 1.96\widehat{\text{se}}$ for 95% coverage. These are based on an assumption of normality for $\hat{\theta}$. The histogram of the 2000 bootstrap replications of $\hat{\theta}$, as seen in Figure 10.2, disabuses belief in even approximate normality. Compared with classical methods, a massive amount of computation has gone into the histogram, but this will pay off in Chapter 11 with more accurate confidence limits. We can claim a double reward here for bootstrap methods: much wider applicability and improved inferences. The bootstrap histogram—invisible to classical statisticians—nicely illustrates the advantages of computer-age statistical inference.

10.3 Resampling Plans

There is a second way to think about the jackknife and the bootstrap: as algorithms that reweight, or *resample*, the original data vector $\boldsymbol{x} =$

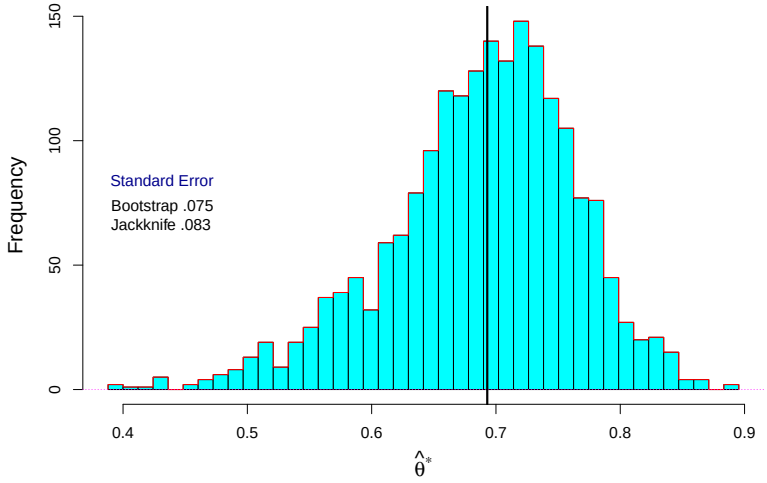

Figure 10.2 Histogram of $B = 2000$ bootstrap replications $\hat{\theta}^*$ for the eigenratio statistic (10.22) for the student score data. The vertical black line is at $\hat{\theta} = .693$. The long left tail shows that normality is a dangerous assumption in this case.

$(x_1, x_2, \ldots, x_n)'$. At the price of a little more abstraction, resampling connects the two algorithms and suggests a class of other possibilities.

A *resampling vector* $\boldsymbol{P} = (P_1, P_2, \ldots, P_n)'$ is by definition a vector of nonnegative weights summing to 1,

$$\boldsymbol{P} = (P_1, P_2, \ldots, P_n)' \qquad \text{with } P_i \geq 0 \text{ and } \sum_{i=1}^{n} P_i = 1. \tag{10.23}$$

That is, $\boldsymbol{P}$ is a member of the simplex $\mathcal{S}_n$ (5.39). Resampling plans operate by holding the original data set $\boldsymbol{x}$ fixed, and seeing how the statistic of interest $\hat{\theta}$ changes as the weight vector $\boldsymbol{P}$ varies across $\mathcal{S}_n$.

We denote the value of $\hat{\theta}$ for a vector putting weight P_i on x_i as

$$\hat{\theta}^* = S(\boldsymbol{P}), \tag{10.24}$$

the star notation now indicating any reweighting, not necessarily from bootstrapping; $\hat{\theta} = s(\boldsymbol{x})$ describes the behavior of $\hat{\theta}$ in the real world (10.17), while $\hat{\theta}^* = S(\boldsymbol{P})$ describes it in the resampling world. For the sample mean $s(\boldsymbol{x}) = \bar{x}$, we have $S(\boldsymbol{P}) = \sum_1^n P_i x_i$. The unbiased estimate of

variance $s(\boldsymbol{x}) = \sum_i^n (x_i - \bar{x})^2/(n-1)$ can be seen to have

$$S(\boldsymbol{P}) = \frac{n}{n-1}\left[\sum_{i=1}^{n} P_i x_i^2 - \left(\sum_{i=1}^{n} P_i x_i\right)^2\right]. \tag{10.25}$$

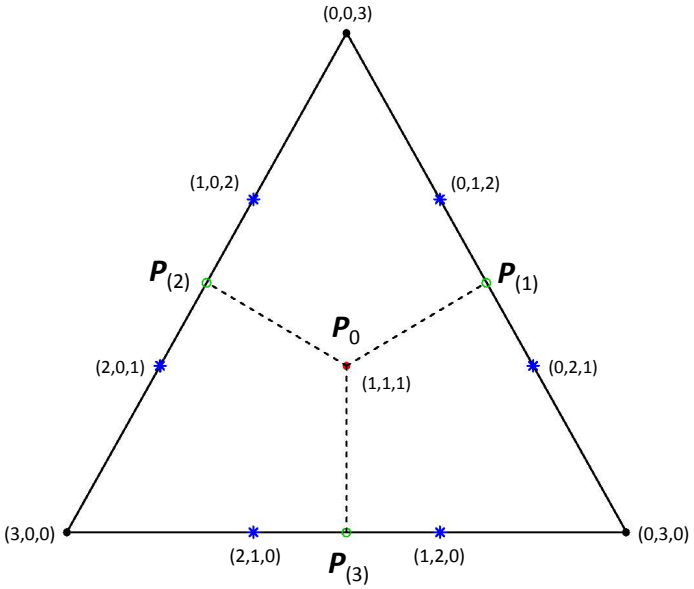

Figure 10.3 Resampling simplex for sample size $n = 3$. The center point is $\boldsymbol{P}_0$ (10.26); the green circles are the jackknife points $\boldsymbol{P}_{(i)}$ (10.28); triples indicate bootstrap resampling numbers (N_1, N_2, N_3) (10.29). The *bootstrap probabilities* are $6/27$ for $\boldsymbol{P}_0$, $1/27$ for each corner point, and $3/27$ for each of the six starred points.

Letting

$$\boldsymbol{P}_0 = (1, 1, \ldots, 1)'/n, \tag{10.26}$$

the resampling vector putting equal weight on each value x_i, we require in the definition of $S(\cdot)$ that

$$S(\boldsymbol{P}_0) = s(\boldsymbol{x}) = \hat{\theta}, \tag{10.27}$$

the original estimate. The ith jackknife value $\hat{\theta}_{(i)}$ (10.5) corresponds to

resampling vector

$$\boldsymbol{P}_{(i)} = (1, 1, \ldots, 1, 0, 1, \ldots, 1)'/(n-1), \tag{10.28}$$

with 0 in the ith place. Figure 10.3 illustrates the resampling simplex $\mathcal{S}_3$ applying to sample size $n = 3$, with the center point being $\boldsymbol{P}_0$ and the open circles the three possible jackknife vectors $\boldsymbol{P}_{(i)}$.

With $n = 3$ sample points $\{x_1, x_2, x_3\}$ there are only 10 distinct bootstrap vectors (10.13), also shown in Figure 10.3. Let

$$N_i = \#\{x_j^* = x_i\}, \tag{10.29}$$

the number of bootstrap draws in $\boldsymbol{x}^*$ equaling x_i. The triples in the figure are (N_1, N_2, N_3), for example $(1, 0, 2)$ for $\boldsymbol{x}^*$ having x_1 once and x_3 twice.[4] The bootstrap resampling vectors are of the form

$$\boldsymbol{P}^* = (N_1, N_2, \ldots, N_n)'/n, \tag{10.30}$$

where the N_i are nonnegative integers summing to n. According to definition (10.13) of bootstrap sampling, the vector $\boldsymbol{N} = (N_1, N_2, \ldots, N_n)'$ follows a multinomial distribution (5.38) with n draws on n equally likely categories,

$$\boldsymbol{N} \sim \text{Mult}_n(n, \boldsymbol{P}_0). \tag{10.31}$$

This gives bootstrap probability (5.37)

$$\frac{n!}{N_1!N_2!\ldots N_n!}\,\frac{1}{n^n} \tag{10.32}$$

on $\boldsymbol{P}^*$ (10.30).

Figure 10.3 is misleading in that the jackknife vectors $\boldsymbol{P}_{(i)}$ appear only slightly closer to $\boldsymbol{P}_0$ than are the bootstrap vectors $\boldsymbol{P}^*$. As n grows large they are, in fact, an order of magnitude closer. Subtracting (10.26) from (10.28) gives Euclidean distance

$$\|\boldsymbol{P}_{(i)} - \boldsymbol{P}_0\| = 1\Big/\sqrt{n(n-1)}. \tag{10.33}$$

For the bootstrap, notice that N_i in (10.29) has a binomial distribution,

$$N_i \sim \text{Bi}\left(n, \frac{1}{n}\right), \tag{10.34}$$

[4] A hidden assumption of definition (10.24) is that $\hat{\theta} = s(\boldsymbol{x})$ has the same value for any permutation of $\boldsymbol{x}$, so for instance $s(x_1, x_3, x_3) = s(x_3, x_1, x_3) = S(1/3, 0, 2/3)$.

with mean 1 and variance $(n-1)/n$. Then $P_i^* = N_i/n$ has mean and variance $(1/n, (n-1)/n^3)$. Adding over the n coordinates gives the expected root mean square distance for bootstrap vector $\boldsymbol{P}^*$,

$$\left(E\|\boldsymbol{P}^* - \boldsymbol{P}_0\|^2\right)^{1/2} = \sqrt{(n-1)/n^2}, \tag{10.35}$$

an order of magnitude $\sqrt{n}$ times further than (10.33).

The function $S(\boldsymbol{P})$ has approximate directional derivative

$$D_i = \frac{S(\boldsymbol{P}_{(i)}) - S(\boldsymbol{P}_0)}{\|\boldsymbol{P}_{(i)} - \boldsymbol{P}_0\|} \tag{10.36}$$

in the direction from $\boldsymbol{P}_0$ toward $\boldsymbol{P}_{(i)}$ (measured along the dashed lines in Figure 10.3). D_i measures the slope of function $S(\boldsymbol{P})$ at $\boldsymbol{P}_0$, in the direction of $\boldsymbol{P}_{(i)}$. Formula (10.12) shows $\widehat{\text{se}}_{\text{jack}}$ as proportional to the root mean square of the slopes.

If $S(\boldsymbol{P})$ is a *linear* function of $\boldsymbol{P}$, as it is for the sample mean, it turns out that $\widehat{\text{se}}_{\text{jack}}$ equals $\widehat{\text{se}}_{\text{boot}}$ (except for the fudge factor $(n-1)/n$ in (10.6)). Most statistics are not linear, and then the local jackknife resamples may provide a poor approximation to the full resampling behavior of $S(\boldsymbol{P})$. This was the case at one point in Figure 10.1.

With only 10 possible resampling points $\boldsymbol{P}^*$, we can easily evaluate the *ideal* bootstrap standard error estimate

$$\widehat{\text{se}}_{\text{boot}} = \left[\sum_{k=1}^{10} p_k \left(\hat{\theta}^{*k} - \hat{\theta}^{*\cdot}\right)^2\right]^{1/2}, \quad \hat{\theta}^{*\cdot} = \sum_{k=1}^{10} p_k \hat{\theta}^{*k}, \tag{10.37}$$

with $\hat{\theta}^{*k} = S(\boldsymbol{P}^k)$ and p_k the probability from (10.32) (listed in Figure 10.3). This rapidly becomes impractical. The number of distinct bootstrap samples for n points turns out to be

$$\binom{2n-1}{n}. \tag{10.38}$$

For $n = 10$ this is already 92,378, while $n = 20$ gives 6.9×10^{10} distinct possible resamples. Choosing B vectors $\boldsymbol{P}^*$ at random, which is what algorithm (10.13)–(10.15) effectively is doing, makes the un-ideal bootstrap standard error estimate (10.16) almost as accurate as (10.37) for B as small as 200 or even less.

The luxury of examining the resampling surface provides a major advantage to modern statisticians, both in inference and methodology. A variety of other resampling schemes have been proposed, a few of which follow.

The Infinitesimal Jackknife

Looking at Figure 10.3 again, the vector

$$\boldsymbol{P}_i(\epsilon) = (1-\epsilon)\boldsymbol{P}_0 + \epsilon\boldsymbol{P}_{(i)} = \boldsymbol{P}_0 + \epsilon(\boldsymbol{P}_{(i)} - \boldsymbol{P}_0) \tag{10.39}$$

lies proportion ϵ of the way from $\boldsymbol{P}_0$ to $\boldsymbol{P}_{(i)}$. Then

$$\tilde{D}_i = \lim_{\epsilon\to 0} \frac{S\left(\boldsymbol{P}_i(\epsilon)\right) - S(\boldsymbol{P}_0)}{\epsilon\|\boldsymbol{P}_{(i)} - \boldsymbol{P}_0\|} \tag{10.40}$$

exactly defines the direction derivative at $\boldsymbol{P}_0$ in the direction of $\boldsymbol{P}_{(i)}$. The infinitesimal jackknife estimate of standard error is

$$\widehat{\text{se}}_{\text{IJ}} = \left(\sum_{i=1}^{n} \tilde{D}_i^2/n^2\right)^{1/2}, \tag{10.41}$$

usually evaluated numerically by setting ϵ to some small value in (10.40)–(10.41) (rather than $\epsilon = 1$ in (10.12)). We will meet the infinitesimal jackknife again in Chapters 17 and 20.

Multisample Bootstrap

The median difference between the **AML** and the **ALL** scores in Figure 1.4 is

$$\texttt{mediff} = 0.968 - 0.733 = 0.235. \tag{10.42}$$

How accurate is 0.235? An appropriate form of bootstrapping draws 25 times with replacement from the 25 **AML** patients, 47 times with replacement from the 47 **ALL** patients, and computes $\texttt{mediff}^*$ as the difference between the medians of the two bootstrap samples. (Drawing one bootstrap sample of size 72 from all the patients would result in random sample sizes for the $\texttt{AML}^*/\texttt{ALL}^*$ groups, adding inappropriate variability to the frequentist standard error estimate.)

A histogram of $B = 500$ $\texttt{mediff}^*$ values appears in Figure 10.4. They give $\widehat{\text{se}}_{\text{boot}} = 0.074$. The estimate (10.42) is 3.18 $\widehat{\text{se}}$ units above zero, agreeing surprisingly well with the usual two-sample t-statistic 3.13 (based on *mean* differences), and its permutation histogram Figure 4.3. Permutation testing can be considered another form of resampling.

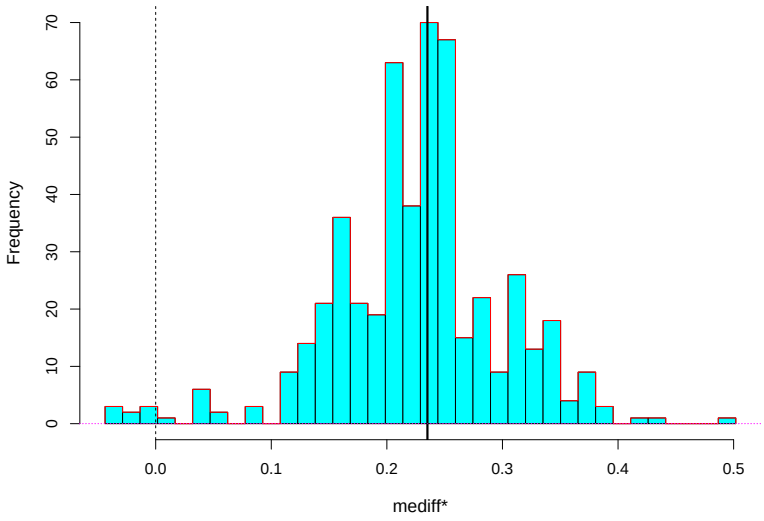

Figure 10.4 $B = 500$ bootstrap replications for the median difference between the `AML` and `ALL` scores in Figure 1.4, giving $\widehat{\text{se}}_{\text{boot}} = 0.074$. The observed value `mediff` $= 0.235$ (vertical black line) is more than 3 standard errors above zero.

Moving Blocks Bootstrap

Suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, instead of being an iid sample (10.2), is a time series. That is, the x values occur in a meaningful order, perhaps with nearby observations highly correlated with each other. Let $\mathcal{B}_m$ be the set of contiguous blocks of length m, for example

$$\mathcal{B}_3 = \{(x_1, x_2, x_3), (x_2, x_3, x_4), \ldots, (x_{n-2}, x_{n-1}, x_n)\}. \tag{10.43}$$

Presumably, m is chosen large enough that correlations between x_i and x_j, $|j - i| > m$, are neglible. The moving block bootstrap first selects n/m blocks from $\mathcal{B}_m$, and assembles them in random order to construct a bootstrap sample $\boldsymbol{x}^*$. Having constructed B such samples, $\widehat{\text{se}}_{\text{boot}}$ is calculated as in (10.15)–(10.16).

The Bayesian Bootstrap

Let $G_1, G_2, \ldots, G_n$ be independent one-sided exponential variates (denoted Gam(1,1) in Table 5.1), each having density $\exp(-x)$ for $x > 0$.

The Bayesian bootstrap uses resampling vectors

$$\boldsymbol{P}^* = (G_1, G_2, \ldots, G_n) \Big/ \sum_1^n G_i. \tag{10.44}$$

It can be shown that $\boldsymbol{P}^*$ is then uniformly distributed over the resampling simplex $\mathcal{S}_n$; for $n = 3$, uniformly distributed over the triangle in Figure 10.3. Prescription (10.44) is motivated by assuming a Jeffreys-style uninformative prior distribution (Section 3.2) on the unknown distribution F (10.2).

Distribution (10.44) for $\boldsymbol{P}^*$ has mean vector and covariance matrix

$$\boldsymbol{P}^* \sim \left[\boldsymbol{P}_0, \frac{1}{n+1}\left(\text{diag}(\boldsymbol{P}_0) - \boldsymbol{P}_0\boldsymbol{P}_0'\right)\right]. \tag{10.45}$$

This is almost identical to the mean and covariance of bootstrap resamples $\boldsymbol{P}^* \sim \text{Mult}_n(n, \boldsymbol{P}_0)/n$,

$$\boldsymbol{P}^* \sim \left[\boldsymbol{P}_0, \frac{1}{n}\left(\text{diag}(\boldsymbol{P}_0) - \boldsymbol{P}_0\boldsymbol{P}_0'\right)\right], \tag{10.46}$$

(5.40). The Bayesian bootstrap and the ordinary bootstrap tend to agree, at least for smoothly defined statistics $\hat{\theta}^* = S(\boldsymbol{P}^*)$.

There was some Bayesian disparagement of the bootstrap when it first appeared because of its blatantly frequentist take on estimation accuracy. And yet connections like (10.45)–(10.46) have continued to pop up, as we will see in Chapter 13.

10.4 The Parametric Bootstrap

In our description (10.18) of bootstrap resampling,

$$\hat{F} \overset{\text{iid}}{\longrightarrow} \boldsymbol{x}^* \longrightarrow \hat{\theta}^*, \tag{10.47}$$

there is no need to insist that $\hat{F}$ be the nonparametric MLE of F. Suppose we are willing to assume that the observed data vector $\boldsymbol{x}$ comes from a *parametric family* $\mathcal{F}$ as in (5.1),

$$\mathcal{F} = \left\{f_\mu(\boldsymbol{x}),\ \mu \in \Omega\right\}. \tag{10.48}$$

Let $\hat{\mu}$ be the MLE of μ. The *bootstrap parametric* resamples from $f_{\hat{\mu}}(\cdot)$,

$$f_{\hat{\mu}} \longrightarrow \boldsymbol{x}^* \longrightarrow \hat{\theta}^*, \tag{10.49}$$

and proceeds as in (10.14)–(10.16) to calculate $\widehat{\text{se}}_{\text{boot}}$.

As an example, suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is an iid sample of size n from a normal distribution,

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1), \qquad i = 1, 2, \ldots, n. \tag{10.50}$$

Then $\hat{\mu} = \bar{x}$, and a parametric bootstrap sample is $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$, where

$$x_i^* \overset{\text{iid}}{\sim} \mathcal{N}(\bar{x}, 1), \qquad i = 1, 2, \ldots, n. \tag{10.51}$$

More adventurously, if $\mathcal{F}$ were a family of time series models for $\boldsymbol{x}$, algorithm (10.49) would still apply (now without any iid structure): $\boldsymbol{x}^*$ would be a time series sampled from model $f_{\hat{\mu}}(\cdot)$, and $\hat{\theta}^* = s(\boldsymbol{x}^*)$ the resampled statistic of interest. B independent realizations $\boldsymbol{x}^{*b}$ would give $\hat{\theta}^{*b}$, $b = 1, 2, \ldots, B$, and $\widehat{\text{se}}_{\text{boot}}$ from (10.16).

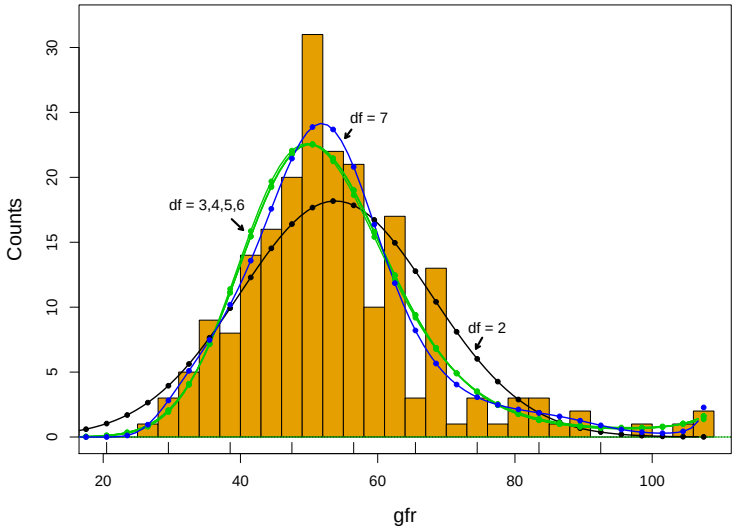

Figure 10.5 The `gfr` data of Figure 5.7 (histogram). Curves show the MLE fits from polynomial Poisson models, for degrees of freedom df $= 2, 3, \ldots, 7$. The points on the curves show the fits computed at the centers $x_{(j)}$ of the bins, with the responses being the counts in the bins. The dashes at the base of the plot show the nine `gfr` values appearing in Table 10.2.

As an example of parametric bootstrapping, Figure 10.5 expands the `gfr` investigation of Figure 5.7. In addition to the seventh-degree polynomial fit (5.62), we now show lower-degree polynomial fits for 2, 3, 4, 5,

and 6 degrees of freedom; df = 2 obviously gives a poor fit; df = 3, 4, 5, 6 give nearly identical curves; df = 7 gives only a slightly better fit to the raw data.

The plotted curves were obtained from the Poisson regression method used in Section 8.3.[5]

- The x-axis was partitioned into $K = 32$ bins, with endpoints 13, 16, 19, ..., 109, and centerpoints, say,

$$\boldsymbol{x}_{(\,)} = (x_{(1)}, x_{(2)}, \ldots, x_{(K)}), \tag{10.52}$$

  $x_{(1)} = 14.5$, $x_{(2)} = 17.5$, etc.
- Count vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)$ was computed

$$y_k = \#\{x_i \text{ in bin}_k\} \tag{10.53}$$

  (so $\boldsymbol{y}$ gives the heights of the bars in Figure 10.5).
- An independent Poisson model was assumed for the counts,

$$y_k \overset{\text{ind}}{\sim} \text{Poi}(\mu_k) \qquad \text{for } k = 1, 2, \ldots, K. \tag{10.54}$$

- The parametric model of degree "df" assumed that the μ_k values were described by an exponential polynomial of degree df in the $x_{(k)}$ values,

$$\log(\mu_k) = \sum_{j=0}^{\text{df}} \beta_j x_{(k)}^j. \tag{10.55}$$

- The MLE $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{\text{df}})$ in model (10.54)–(10.55) was found.[6]
- The plotted curves in Figure 10.5 trace the MLE values $\hat{\mu}_k$,

$$\log(\hat{\mu}_k) = \sum_{j=0}^{\text{df}} \hat{\beta}_j x_{(k)}^j. \tag{10.56}$$

How accurate are the curves? Parametric bootstraps were used to assess their standard errors. That is, Poisson resamples were generated according to

$$y_k^* \overset{\text{ind}}{\sim} \text{Poi}(\hat{\mu}_k) \qquad \text{for } k = 1, 2, \ldots, K, \tag{10.57}$$

and bootstrap MLE values $\hat{\mu}_k^*$ calculated as above, but now based on count vector $\boldsymbol{y}^*$ rather than $\boldsymbol{y}$. All of this was done $B = 200$ times, yielding bootstrap standard errors (10.16).

[5] "Lindsey's method," discussed further in Chapter 15.

[6] A single R command, `glm(y~poly(x,df),family=poisson)` accomplishes this.

Table 10.2 *Bootstrap estimates of standard error for the* `gfr` *density. Poisson regression models* (10.54)–(10.55)*,* df $= 2, 3, \ldots, 7$*, as in Figure 10.5; each* $B = 200$ *bootstrap replications; nonparametric standard errors based on binomial bin counts.*

	Degrees of freedom						Nonparametric standard error
`gfr`	2	3	4	5	6	7	
20.5	.28	.07	.13	.13	.12	.05	.00
29.5	.65	.57	.57	.66	.74	1.11	1.72
38.5	1.05	1.39	1.33	1.52	1.72	1.73	2.77
47.5	1.47	1.91	2.12	1.93	2.15	2.39	4.25
56.5	1.57	1.60	1.79	1.93	1.87	2.28	4.35
65.5	1.15	1.10	1.07	1.31	1.34	1.27	1.72
74.5	.76	.61	.62	.68	.81	.71	1.72
83.5	.40	.30	.40	.38	.49	.68	1.72
92.5	.13	.20	.29	.29	.34	.46	.00

The results appear in Table 10.2, showing $\widehat{\text{se}}_{\text{boot}}$ for df $= 2, 3, \ldots, 7$ degrees of freedom evaluated at nine values of `gfr`. Variability generally increases with increasing df, as expected. Choosing a "best" model is a compromise between standard error and possible definitional bias as suggested by Figure 10.5, with perhaps df $= 3$ or 4, the winner.

If we kept increasing the degrees of freedom, eventually (at df $= 32$) we would exactly match the bar heights y_k in the histogram. At this point the parametric bootstrap would merge into the nonparametric bootstrap. "Nonparametric" is another name for "very highly parameterized." The huge sample sizes associated with modern applications have encouraged nonparametric methods, on the sometimes mistaken ground that estimation efficiency is no longer of concern. It is costly here, as the "nonparametric" column of Table 10.2 shows.[7]

Figure 10.6 returns to the student score eigenratio calculations of Figure 10.2. The solid histogram shows 2000 parametric bootstrap replications (10.49), with $f_{\hat{\mu}}$ the five-dimensional bivariate normal distribution $\mathcal{N}_5(\bar{x}, \hat{\Sigma})$. Here $\bar{x}$ and $\hat{\Sigma}$ are the usual MLE estimates for the expectation vector and covariance matrix based on the 22 five-component student score vectors. It is narrower than the corresponding nonparametric bootstrap histogram, with $\widehat{\text{se}}_{\text{boot}} = 0.070$ compared with the nonparametric estimate

[7] These are the binomial standard errors $[y_k(1 - y_k)/n]^{1/2}$, $n = 211$. The nonparametric results look much more competitive when estimating cdf's rather than densities.

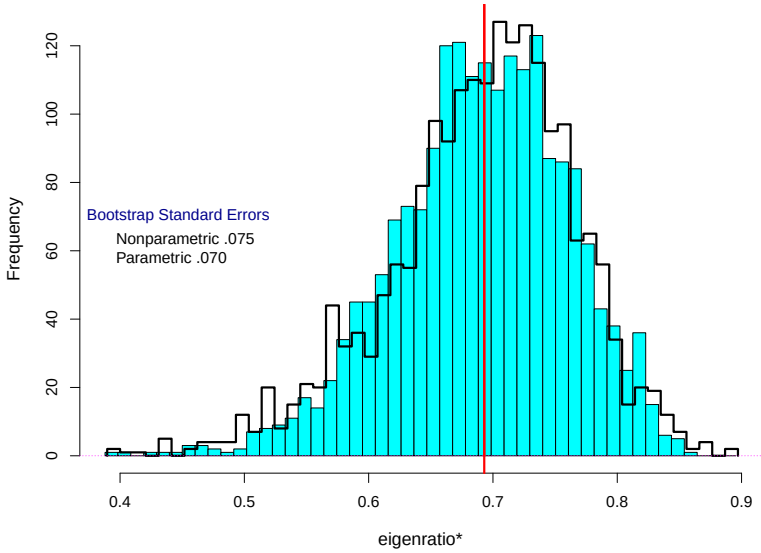

Figure 10.6 Eigenratio example, student score data. *Solid histogram* $B = 2000$ parametric bootstrap replications $\hat{\theta}^*$ from the five-dimensional normal MLE; *line histogram* the 2000 nonparametric replications of Figure 10.2. MLE $\hat{\theta} = .693$ is vertical red line.

0.075. (Note the different histogram bin limits from Figure 10.2, changing the details of the nonparametric histogram.)

Parametric families act as *regularizers*, smoothing out the raw data and de-emphasizing outliers. In fact the student score data is not a good candidate for normal modeling, having at least one notable outlier,[8] casting doubt on the smaller estimate of standard error.

The classical statistician could only imagine a mathematical device that given any statistic $\hat{\theta} = s(\boldsymbol{x})$ would produce a formula for its standard error, as formula (1.2) does for $\bar{x}$. The electronic computer *is* such a device. As harnessed by the bootstrap, it automatically produces a numerical estimate of standard error (though not a formula), with no further cleverness required. Chapter 11 discusses a more ambitious substitution of computer power for mathematical analysis: the bootstrap computation of confidence intervals.

[8] As revealed by examining scatterplots of the five variates taken two at a time. Fast and painless plotting is another advantage for twenty-first-century data analysts.

10.5 Influence Functions and Robust Estimation

The sample mean played a dominant role in classical statistics for reasons heavily weighted toward mathematical tractibility. Beginning in the 1960s, an important counter-movement, *robust estimation*, aimed to improve upon the statistical properties of the mean. A central element of that theory, the *influence function*, is closely related to the jackknife and infinitesimal jackknife estimates of standard error.

We will only consider the case where $\mathcal{X}$, the sample space, is an interval of the real line. The unknown probability distribution F yielding the iid sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ in (10.2) is now the cdf of a density function $f(x)$ on $\mathcal{X}$. A parameter of interest, i.e., a function of F, is to be estimated by the plug-in principle, $\hat{\theta} = T(\hat{F})$, where, as in Section 10.2, $\hat{F}$ is the empirical probability distribution putting probability $1/n$ on each sample point x_i. For the mean,

$$\theta = T(F) = \int_{\mathcal{X}} x f(x)\, dx \quad \text{and} \quad \hat{\theta} = T\left(\hat{F}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{10.58}$$

(In Riemann–Stieltjes notation, $\theta = \int x dF(x)$ and $\hat{\theta} = \int x d\hat{F}(x)$.)

The influence function of $T(F)$, evaluated at point x in $\mathcal{X}$, is defined to be

$$\mathrm{IF}(x) = \lim_{\epsilon \to 0} \frac{T\left((1-\epsilon)F + \epsilon\delta_x\right) - T(F)}{\epsilon}, \tag{10.59}$$

where δ_x is the "one-point probability distribution" putting probability 1 on x. In words, $\mathrm{IF}(x)$ measures the differential effect of modifying F by putting additional probability on x. For the mean $\theta = \int x f(x) dx$ we calculate that

$$\mathrm{IF}(x) = x - \theta. \tag{10.60}$$

†5 A fundamental theorem† says that $\hat{\theta} = T(\hat{F})$ is approximately

$$\hat{\theta} \doteq \theta + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}(x_i), \tag{10.61}$$

with the approximation becoming exact as n goes to infinity. This implies that $\hat{\theta} - \theta$ is, approximately, the mean of the n iid variates $\mathrm{IF}(x_i)$, and that the variance of $\hat{\theta}$ is approximately

$$\mathrm{var}\left\{\hat{\theta}\right\} \doteq \frac{1}{n}\,\mathrm{var}\left\{\mathrm{IF}(x)\right\}, \tag{10.62}$$

var{IF(x)} being the variance of IF(x) for any one draw of x from F. For the sample mean, using (10.60) in (10.62) gives the familiar equality

$$\text{var}\{\bar{x}\} = \frac{1}{n}\,\text{var}\{x\}. \tag{10.63}$$

The sample mean suffers from an *unbounded* influence function (10.60), which grows ever larger as x moves farther from θ. This makes $\bar{x}$ unstable against heavy-tailed densities such as the Cauchy (4.39). Robust estimation theory seeks estimators $\hat{\theta}$ of bounded influence, that do well against heavy-tailed densities without giving up too much efficiency against light-tailed densities such as the normal. Of particular interest have been the trimmed mean and its close cousin the winsorized mean.

Let $x^{(\alpha)}$ denote the 100αth percentile of distribution F, satisfying $F(x^{(\alpha)}) = \alpha$ or equivalently

$$\alpha = \int_{-\infty}^{x^{(\alpha)}} f(x)\,dx. \tag{10.64}$$

The *αth trimmed mean of F*, $\theta_{\text{trim}}(\alpha)$, is defined as

$$\theta_{\text{trim}}(\alpha) = \frac{1}{1-2\alpha}\int_{x^{(\alpha)}}^{x^{(1-\alpha)}} xf(x)\,dx, \tag{10.65}$$

the mean of the central $1-2\alpha$ portion of F, trimming off the lower and upper α portions. This is not the same as the *αth winsorized mean* $\theta_{\text{wins}}(\alpha)$,

$$\theta_{\text{wins}}(\alpha) = \int_{\mathcal{X}} W(x)f(x)\,dx, \tag{10.66}$$

where

$$W(x) = \begin{cases} x^{(\alpha)} & \text{if } x \le x^{(\alpha)} \\ x & \text{if } x^{(\alpha)} \le x \le x^{(1-\alpha)} \\ x^{(1-\alpha)} & \text{if } x \ge x^{(1-\alpha)}; \end{cases} \tag{10.67}$$

$\theta_{\text{trim}}(\alpha)$ removes the outer portions of F, while $\theta_{\text{wins}}(\alpha)$ moves them into $x^{(\alpha)}$ or $x^{(1-\alpha)}$. In practice, empirical versions $\hat{\theta}_{\text{trim}}(\alpha)$ and $\hat{\theta}_{\text{wins}}(\alpha)$ are used, substituting the empirical density $\hat{f}$, with probability $1/n$ at each x_i, for f.

There turns out to be an interesting relationship between the two: the influence function of $\theta_{\text{trim}}(\alpha)$ is a function of $\theta_{\text{wins}}(\alpha)$,

$$\text{IF}_\alpha(x) = \frac{W(x) - \theta_{\text{wins}}(\alpha)}{1-2\alpha}. \tag{10.68}$$

This is pictured in Figure 10.7, where we have plotted empirical influence

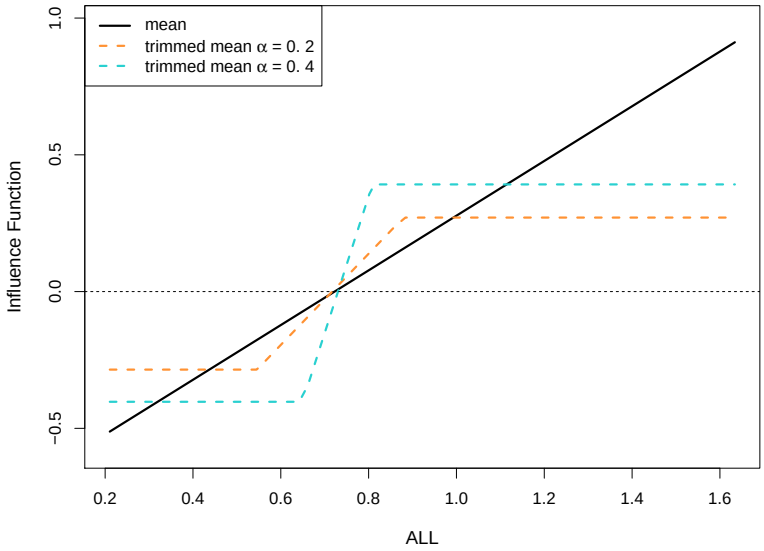

Figure 10.7 Empirical influence functions for the 47 leukemia **ALL** scores of Figure 1.4. The two dashed curves are $\mathrm{IF}_\alpha(x)$ for the trimmed means (10.68), for $\alpha = 0.2$ and $\alpha = 0.4$. The solid curve is $\mathrm{IF}(x)$ for the sample mean $\bar{x}$ (10.60).

functions (plugging in $\hat{F}$ for F in definition (10.59)) relating to the 47 leukemia **ALL** scores of Figure 1.4: $\mathrm{IF}_{0.2}(x)$ and $\mathrm{IF}_{0.4}(x)$ are plotted, along with $\mathrm{IF}_0(x)$ (10.60), that is, for the mean.

Table 10.3 *Trimmed means and their bootstrap standard deviations for the 47 leukemia* **ALL** *scores of Figure 1.4;* $B = 1000$ *bootstrap replications for each trim value. The last column gives empirical influence function estimates of the standard error, which are also the infinitesimal jackknife estimates* (10.41)*. These fail for the median.*

	Trim	Trimmed mean	Bootstrap sd	(IFse)
Mean	.0	.752	.040	(.040)
	.1	.729	.038	(.034)
	.2	.720	.035	(.034)
	.3	.725	.044	(.044)
	.4	.734	.047	(.054)
Median	.5	.733	.053	

The upper panel of Figure 1.4 shows a moderately heavy right tail for the **ALL** distribution. Would it be more efficient to estimate the center of the distribution with a trimmed mean rather than $\bar{x}$? The bootstrap provides an answer: $\widehat{\text{se}}_{\text{boot}}$ (10.16) was calculated for $\bar{x}$ and $\hat{\theta}_{\text{trim}}(\alpha)$, $\alpha = 0.1, 0.2, 0.3, 0.4$, and 0.5, the last being the sample median. It appears that $\hat{\theta}_{\text{trim}}(0.2)$ is moderately better than $\bar{x}$. This brings up an important question discussed in Chapter 20: if we use something like Table 10.3 to select an estimator, how does the selection process affect the accuracy of the resulting estimate?

We might also use the square root of formula (10.62) to estimate the standard errors of the various estimators, plugging in the empirical influence function for $\text{IF}(x)$. This turns out to be the same as using the infinitesimal jackknife (10.41). These appear in the last column of Table 10.3. Predictably, this approach fails for the sample median, whose influence function is a square wave, sharply discontinuous at the median θ,

$$\text{IF}(x) = \pm 1 \big/ (2f(\theta)) . \tag{10.69}$$

Robust estimation offers a nice illustration of statistical progress in the computer age. Trimmed means go far back into the classical era. Influence functions are an insightful inferential tool for understanding the tradeoffs in trimmed mean estimation. And finally the bootstrap allows easy assessment of the accuracy of robust estimation, including some more elaborate ones not discussed here.

10.6 Notes and Details

Quenouille (1956) introduced what is now called the jackknife estimate of bias. Tukey (1958) realized that Quenouille-type calculations could be repurposed for nonparametric standard-error estimation, inventing formula (10.6) and naming it "the jackknife," as a rough and ready tool. Miller's important 1964 paper, "A trustworthy jackknife," asked when formula (10.6) could be trusted. (Not for the median.)

The bootstrap (Efron, 1979) began as an attempt to better understand the jackknife's successes and failures. Its name celebrates Baron Munchausen's success in pulling himself up by his own bootstraps from the bottom of a lake. Burgeoning computer power soon overcame the bootstrap's main drawback, prodigous amounts of calculation, propelling it into general use. Meanwhile, 1000+ theoretical papers were published asking when the bootstrap itself could be trusted. (Most but not all of the time in common practice).

A main reference for the chapter is Efron's 1982 monograph *The Jackknife, the Bootstrap and Other Resampling Plans*. Its Chapter 6 shows the equality of three nonparametric standard error estimates: Jaeckel's (1972) infinitesimal jackknife (10.41); the empirical influence function estimate, based on (10.62); and what is known as the nonparametric delta method.

Bootstrap Packages

Various bootstrap packages in **R** are available on the CRAN contributed-packages web site, **bootstrap** being an ambitious one. Algorithm 10.1 shows a simple **R** program for nonparametric bootstrapping. Aside from bookkeeping, it's only a few lines long.

Algorithm 10.1 *An R program for the nonparametric bootstrap.*

```
Boot <- function (x, B, func, ...){
 # x is data vector or matrix (with each row a case)
 # B is number of bootstrap replications
 # func is R function that inputs a data vector or
 # matrix and returns a numeric number or vector
 # ... other arguments for func
    x <- as.matrix(x)
    n <- nrow(x)
    f0=func(x,...) # get size of output
    fmat <- matrix(0,length(f0),B)
    for (b in 1:B) {
         i=sample(1:n, n, replace = TRUE)
         fmat[,b] <- func(x[i, ],...)
     }
    drop(fmat)
}
```

†1 [p. 158] *The jackknife standard error.* The 1982 monograph also contains Efron and Stein's (1981) result on the bias of the jackknife variance estimate, the square of formula (10.6): modulo certain sample size considerations, the expectation of the jackknife variance estimate is biased upward for the true variance.

For the sample mean $\bar{x}$, the jackknife yields exactly the usual variance estimate (1.2), $\sum_i (x_i - \bar{x})^2/(n(n-1))$, while the ideal bootstrap estimate ($B \to \infty$) gives

$$\sum_{i=1}^{n} (x_i - \bar{x})^2/n^2. \tag{10.70}$$

As with the jackknife, we could append a fudge factor to get perfect agreement with (1.2), but there is no real gain in doing so.

†2 [p. 161] *Bootstrap sample sizes.* Let $\widehat{\text{se}}_B$ indicate the bootstrap standard error estimate (10.16) based on B replications, and $\widehat{\text{se}}_\infty$ the "ideal bootstrap," $B \to \infty$. In any actual application, there are diminishing returns from increasing B past a certain point, because $\widehat{\text{se}}_\infty$ is itself a statistic whose value varies with the observed sample $\boldsymbol{x}$ (as in (10.70)), leaving an irreducible remainder of randomness in any standard error estimate. Section 6.4 of Efron and Tibshirani (1993) shows that $B = 200$ will almost always be plenty (for standard errors, but not for bootstrap confidence intervals, Chapter 11). Smaller numbers, 25 or even less, can still be quite useful in complicated situations where resampling is expensive. An early complaint, "Bootstrap estimates are random," is less often heard in an era of frequent and massive simulations.

†3 [p. 161] *The Bayesian bootstrap.* Rubin (1981) suggested the Bayesian bootstrap (10.44). Section 10.6 of Efron (1982) used (10.45)–(10.46) as an objective Bayes justification for what we will call the percentile-method bootstrap confidence intervals in Chapter 12.

†4 [p. 161] *Jackknife-after-bootstrap.* For the eigenratio example displayed in Figure 10.2, $B = 2000$ nonparametric bootstrap replications gave $\widehat{\text{se}}_{\text{boot}} = 0.075$. How accurate is this value? Bootstrapping the bootstrap seems like too much work, perhaps 200 times 2000 resamples. It turns out, though, that we can use the jackknife to estimate the variability of $\widehat{\text{se}}_{\text{boot}}$ based on just the original 2000 replications.

Now the deleted sample estimate in (10.6) is $\widehat{\text{se}}_{\text{boot}(i)}$. The key idea is to consider those bootstrap samples $\boldsymbol{x}^*$ (10.13), among the original 2000, that *do not include the point* x_i. About 37% of the original B samples will be in this subset. Section 19.4 of Efron and Tibshirani (1993) shows that applying definition (10.16) to this subset gives $\widehat{\text{se}}_{\text{boot}(i)}$. For the estimate of Figure 10.2, the jackknife-after-bootstrap calculations gave $\widehat{\text{se}}_{\text{jack}} = 0.022$ for $\widehat{\text{se}}_{\text{boot}} = 0.075$. In other words, 0.075 isn't very accurate, which is to be expected for the standard error of a complicated statistic estimated from only $n = 22$ observations. An infinitesimal jackknife version of this technique will play a major role in Chapter 20.

†5 [p. 174] *A fundamental theorem.* Tukey can justly be considered the founding father of robust statistics, his 1960 paper being especially influential. Huber's celebrated 1964 paper brought the subject into the realm of high-concept mathematical statistics. *Robust Statistics: The Approach Based on Influence Functions*, the 1986 book by Hampel *et al.*, conveys the breadth of a subject only lightly scratched in our Section 10.5. Hampel (1974)

introduced the influence function as a statistical tool. Boos and Serfling (1980) verified expression (10.62). Qualitative notions of robustness, more than specific theoretical results, have had a continuing influence on modern data analysis.