SIAM REVIEW
Vol. 21, No. 4, October 1979

0036-1445/79/2104-0002$01.00/0

COMPUTERS AND THE THEORY OF STATISTICS: THINKING THE UNTHINKABLE*

BRADLEY EFRON†

Abstract. This is a survey article concerning recent advances in certain areas of statistical theory, written for a mathematical audience with no background in statistics. The topics are chosen to illustrate a special point: how the advent of the high-speed computer has affected the development of statistical theory. The topics discussed include nonparametric methods, the jackknife, the bootstrap, cross-validation, error-rate estimation in discriminant analysis, robust estimation, the influence function, censored data, the EM algorithm, and Cox's likelihood function. The exposition is mainly by example, with only a little offered in the way of theoretical development.

1. Introduction. The editors have been kind enough to invite a survey article concerning what's new in the theory of statistics. Any answer to this question must be either incomplete or bewildering to the reader. Here I have tried to be incomplete, selecting my topics to illustrate a special point: how the advent of the high-speed computer has affected the theoretical structure of statistics.

Statistics concerns the comparison of sets of numbers—with each other, with theoretical models, and with past experience. The prototypical scientific question, "Is method A better than method B?," may boil down to the statistical question, "Is set of numbers A bigger than set of numbers B?" If, for example,[1] $A = \{94, 197, 16, 38, 99, 141, 23\}$ and $B = \{52, 104, 146, 10, 50, 31, 40, 27, 46\}$, how can we precisely phrase such a question, in particular the crucial concept of "bigger," and answer it in a scientifically meaningful way? The statistician's standard answer, before 1950, would have been

1) Compute the *t-statistic*, which is the difference between the average of the set A numbers and the average of the set B numbers, divided by a certain quadratic function of all 16 numbers. (The divisor scales the difference between the two averages so that a single table can be used at step 2 below.)

2) Compare the observed value of t with its theoretical distribution calculated under the assumption that all 16 numbers were independently drawn from the same normal ("Gaussian") distribution. This theoretical distribution is published in a standard *t-table*.

3) Decide that set A is really bigger than set B, and not just accidentally bigger, if the observed value of t is in the upper 5% of the theoretical distribution.

The most obvious defect of this procedure is the use of normal distribution theory to determine the critical value at which the observed t becomes "significant." *Nonparametric statistics*, mainly developed since 1950, gives an answer that does not depend upon normal theory:

1) Combine all 16 numbers into one set $C = \{94, 197, \cdots, 46\}$, and consider all 11,440 ways $(=16!/7!9!)$ of partitioning C into two sets "a" and "b," a having 7 members and b having 9 members.

2) For each such partition compute the difference between the average of the set a numbers and the average of the set b numbers, say $\bar{x}_a - \bar{x}_b$. There are 11,440 such differences, one of which is the difference $\bar{x}_A - \bar{x}_B$ corresponding to the data actually observed.

* Received by the editors June 28, 1978, and in revised version December 14, 1978. The preparation of this invited manuscript was supported by the U.S. Army Research Office under Contract DAAG29-79-C-0014.

† Department of Statistics, Stanford University, Stanford, California 94305.

[1] These numbers are cell counts, in thousands, from an experiment involving 16 mice. The 7 mice in set A received an inoculation expected to increase the cell count. The 9 mice in set B did not receive an inoculation.

3) Decide that set A really is bigger than set B if $\bar{x}_A - \bar{x}_B$ is in the upper 5% of the 11,440 $\bar{x}_a - \bar{x}_b$ values.

The nonparametric method pays a stiff computational price for its freedom from normal distribution theory. There is no "significance table," corresponding to the t-table, with which one can compare the observed value of $\bar{x}_A - \bar{x}_B$. Essentially, such a table must be constructed anew for each set of data.[2] On the other hand, more than just freedom from normality assumptions is gained. If a different table has to be constructed for each data set, the statistician may very well choose to table something other than the difference of the averages (which was chosen in the first place because of theoretical properties peculiar to the normal distribution). The recipe for a nonparametric test given above works just as well for the difference of the medians as for the difference of the averages. Or the statistician may first make a nonlinear transformation on each of the 16 numbers, say $y = g(x)$, and compare $\bar{y}_A - \bar{y}_B$ with the tabled values $\bar{y}_a - \bar{y}_b$. Or he may try several different transformations, and several different measures of difference between the two sets of numbers, going through the nonparametric recipe each time, in an attempt to understand how robust the perceived difference between A and B is to changes in the statistical procedure.

The "unthinkable" mentioned in the title is simply the thought that one might be willing to perform 500,000 numerical operations in the analysis of 16 data points. Or one might be willing to perform a billion operations to analyze 500 numbers. Such statements would have seemed insane thirty years ago, when a slow and noisy fifty pound desk calculator which added, subtracted, multiplied, and divided was the most sophisticated computational aid available to most scientists. Most of the statistical theory in common use was developed under the constraint of slow and expensive computation. Now computation is fast and cheap. It is not surprising that new theory is being developed, which takes advantage of the high-speed computer. This paper consists of several examples of such theory, presented, hopefully, in a manner accessible to nonstatisticians.

The set of examples presented here in no way exhausts the range of interesting current work in statistics, not even within the limited context of this article. Some notable omissions include the design of experiments, computer graphics and descriptive statistics ("data analysis"), time series and stochastic processes, Bayes and empirical Bayes methods, Stein estimation and ridge regression, analysis of categorical data, and Monte Carlo methods.[3] Also unmentioned is the vigorous development of numerical analysis methods appropriate to large statistical analyses, see for example Golub and Styan [11], which could easily occupy an article of equal length.

This paper is intended for nonstatisticians, and in order to make it easily readable most of the examples involve artificially small data sets. This belies an important effect of the computer upon statistical thinking. Statistical problems have gotten much bigger, in raw size, during the past 30 years as scientists, emboldened by the data handling capabilities of the computer, have collected larger and larger data sets. It is not unusual these days to work with sets of a million or more numbers, sometimes fitting models which involve thousands of parameters. Even the most timeworn statistical technique,

[2] Shortcuts and approximations are possible, the simplest of which results in using exactly the standard t method described first! R. A. Fisher, the principal figure in the development of normal theory methods, advocated what we have called the nonparametric approach as early as 1935, but most of the theoretical development took place after 1950.

[3] A referee points out that Monte Carlo allows one to go much farther in studying standard statistical methods, such as the t test, under nonstandard (i.e. nonnormal) conditions. This is another way in which the computer impacts on statistical theory.

such as the standard linear model, takes on qualitatively new aspects when applied under these circumstances. A brief discussion of this point can be found in § 8 of Efron [9].

The exposition proceeds by a series of examples, with only an occasional hint of the deeper theoretical questions lurking behind the methods. The references have been chosen for readability as well as importance, and are recommended to readers with some statistical background, who wish to pursue these subjects further.

2. The jackknife. The jackknife,[4] introduced by Quenouille and Tukey in the late 1950's, is an intriguing attempt to solve an important statistical problem: having computed an estimate of some quantity of interest, say a mean or a probability or a correlation, what accuracy can be attached to the estimate? *Accuracy* here refers to the "± something" which often accompanies statistical estimates. The usual ± quantities are based on normal distribution theory, or occasionally some other parametric theory, while the jackknife is a nonparametric technique which makes no such assumptions. Miller [18] gives an excellent review of the subject. Here the explanation will be given in terms of a simple example.

Table 1 refers to the 1973 entering classes of 15 American law schools. For each school two numbers are given,

x_i = average LSAT score of entering students in law school i

y_i = average GPA of entering students in law school i,

TABLE 1
The average LSAT *score and undergraduate* GPA *at* 15 *American law schools, entering classes of* 1973.

School #	1	2	3	4	5	6	7	8
LSAT	576	635	558	578	666	580	555	661
GPA	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
School #	9	10	11	12	13	14	15	
LSAT	651	605	653	575	545	572	594	
GPA	3.36	3.13	3.12	2.74	2.76	2.88	2.96	

$i = 1, 2, \cdots, 15$. (The LSAT is a national test, similar to the Graduate Record Exam, while GPA refers to undergraduate grade point average.) These data are abstracted from Rubin [20]. The data are plotted in Fig. 1.

The correlation coefficient is a measure of association between two sets of numbers, or, in its abstract form, between two infinitely large sets of numbers, usually thought of as two related probability distributions. By definition, the correlation coefficient between the n pairs of numbers (x_i, y_i), $i = 1, 2, \cdots, n$, is

$$\hat{\rho} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \qquad \left(\bar{x} = \sum_{1}^{n} x_i/n, \quad \bar{y} = \sum_{i=1}^{n} y_i/n\right). \tag{2.1}$$

Because of the Cauchy–Schwarz inequality it is always true that $-1 \leqq \hat{\rho} \leqq 1$. The case $\hat{\rho} = 1$ occurs when the (x_i, y_i) pairs lie on a single straight line with positive slope, while

[4] The name "jackknife," coined by Tukey, is meant to convey the notion of a rough and ready tool, useful in a wide variety of situations.

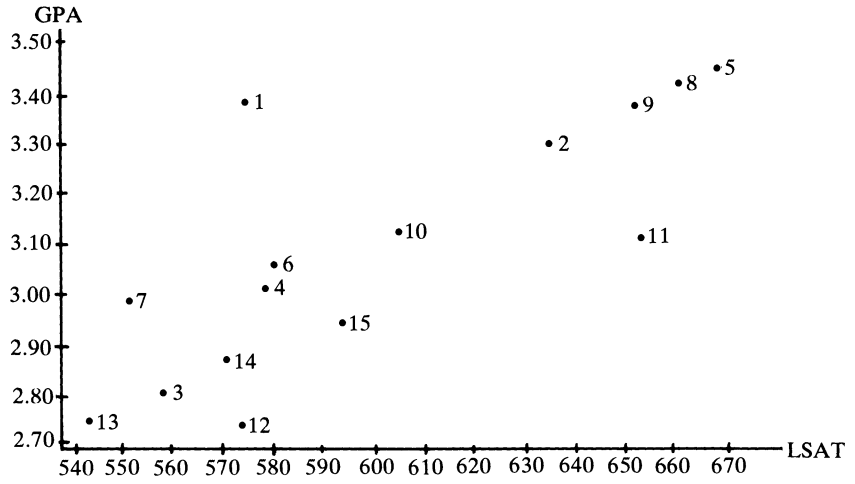

FIG. 1. *A plot of the law school data given in Table* 1.

$\hat{\rho} = -1$ indicates a perfect straight line relationship with negative slope. Figure 1 shows that while the law school data do not go to either of these extremes, they are "positively correlated," i.e. closer to $\hat{\rho} = 1$ than $\hat{\rho} = -1$. The actual value is $\hat{\rho} = .776$, which in most sociological studies would be taken to indicate a strongly positive relationship between the two variables. In plain language, higher LSAT usually goes with higher GPA, and vice versa.

We wish to know how accurate is the estimate $\hat{\rho} = .776$. In asking this question we assume that there is a true correlation ρ which $\hat{\rho}$ is attempting to measure, and which $\hat{\rho}$ would approach if the number of data pairs was increased from $n = 15$ toward $n = \infty$. The most commonly used measure of accuracy is the *standard deviation*,

$$\sigma = \sqrt{E[(\hat{\rho} - \rho)^2]}, \tag{2.2}$$

the root mean square difference of $\hat{\rho}$, based on $n = 15$ pairs, from ρ. Calling (2.2) the standard deviation assumes that $\hat{\rho}$ is unbiased for ρ, that is $E\hat{\rho} = \rho$. This isn't exactly true, but the bias is small enough to be ignored in the law school example, for the sake of simplified presentation. The jackknife theory actually includes a bias correction method which won't be discussed here.

The jackknife estimate of σ, say $\hat{\sigma}^{(J)}$, is obtained by the following procedure:

1) Delete pair (x_i, y_i) from the data set and recompute the correlation coefficient for the remaining 14 pairs. Call this recomputed value $\hat{\rho}_{(i)}$, $i = 1, 2, \cdots, n = 15$.

2) Estimate σ by[5]

$$\hat{\sigma}^{(J)} = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\rho}_{(i)} - \hat{\rho})^2}. \tag{2.3}$$

(It is usual to replace $\hat{\rho}$ by $\sum_{i=1}^{n} \hat{\rho}_{(i)}/n$ in (2.3), again for reasons of bias correction, but the difference in the estimate of $\hat{\sigma}^{(J)}$ is less than .01% in our example.)

[5] Suppose that instead of the correlation coefficient, we wish to estimate the standard deviation of the mean $\bar{x}$ of n numbers $x_1, x_2, \cdots, x_n$. The jackknife procedure, applied to this situation, gives the usual estimate $[\Sigma(x_i - \bar{x})^2/(n(n-1))]^{1/2}$. The factor $(n-1)/n$ in (2.3) is included in order to make the jackknife give this, the "right" answer, for the standard deviation of $\bar{x}$.

Table 2 displays the values of $\hat{\rho}_{(i)} - \hat{\rho}$ for the law school data. The jackknife estimate of accuracy is

$$\hat{\sigma}^{(J)} = \sqrt{.0203} = .143.$$

Notice that we have had to about n times as much computation to get $\hat{\sigma}^{(J)}$ as to get the estimate $\hat{\rho}$ itself.

TABLE 2
The values of $\hat{\rho}_{(i)} - \hat{\rho}$ for the law school data.

i	1	2	3	4	5	6	7	8
$\hat{\rho}_{(i)} - \hat{\rho}$	.116	−.013	−.021	−.000	−.045	.004	.008	−.040
i	9	10	11	12	13	14	15	
$\hat{\rho}_{(i)} - \hat{\rho}$	−.025	−.000	.042	.009	−.036	−.009	.003	

The statistician might now report $\rho = .776 \pm .143$. This means that his best guess of the unknown true value ρ is $\hat{\rho} = .776$, with an expected root mean square error of .143 for $\hat{\rho} - \rho$. If $\hat{\rho} - \rho$ has roughly a normal distribution, which for large amounts of data will always be the case, then the accuracy statement can also be interpreted as

$$\text{Prob}\{\rho \in [.776 - .143,\ .776 + .143]\} \approx .68. \tag{2.4}$$

(Statement (2.4) is based upon the fact that a normal distribution puts 68% of its probability within one standard deviation of the mean.) Interval statements of accuracy like (2.4) have more intuitive appeal than root mean square error.

How good is the estimate $\hat{\sigma}^{(J)}$? We could, if we wanted to, jackknife the entire procedure which computed $\hat{\sigma}^{(J)}$, that is do a second order jackknife, to estimate a standard deviation of $\hat{\sigma}^{(J)}$. (This would require about n^2 times as many calculations as for $\hat{\rho}$.) Instead, we will compare $\hat{\sigma}^{(J)}$ with the traditional normal-theory estimate of $\hat{\rho}$'s standard deviation. In the next section we will calculate the standard deviation in another way which clarifies the connection between the two answers.

Suppose the $n = 15$ pairs (x_i, y_i) are actually drawn from a bivariate normal distribution with correlation coefficient ρ. Then the exact density function of $\hat{\rho}$ can be calculated theoretically. This density function depends only upon ρ, not on the means or standard deviations of x and y, and so can be denoted $f_\rho(\hat{\rho})$; by definition $\int_a^b f_\rho(\hat{\rho})\, d\hat{\rho} = \text{Prob}\{a \leqq \hat{\rho} \leqq b\}$. Figure 2 shows $f_\rho(\cdot)$ for $\rho = .776$, the observed value in the law school samples. It is denoted $f_{\hat{\rho}}(\hat{\rho}^*)$ to preserve the definition of $\hat{\rho}$ as the observed value; $\hat{\rho}^*$ is just a convenient name for the dummy variable in $f_{\hat{\rho}}(\cdot)$. The abscissa is plotted in $\hat{\rho}^* - \hat{\rho}$ to emphasize the deviations of $\hat{\rho}^*$ from $\hat{\rho}$.

We see that the density function is not exactly normal, having a longer tail to the left than to the right, and also is not centered exactly at 0, i.e. at $\hat{\rho}^* = \hat{\rho}$, having instead median value .011. (The normality can be dramatically improved by making Fisher's $\tanh^{-1}$ transformation; see Cramér [5, p. 399].) The traditional normal theory estimate of σ can be described, at the expense of a slight oversimplification, in terms similar to (2.4): look at the central 68% of the distribution described by $f_{\hat{\rho}}(\cdot)$, that is the interval from the 16th percentile to the 84th percentile. Half of the length of this interval is a reasonable definition of the normal-theory estimate of σ, say $\hat{\sigma}^{(N)}$. For $\hat{\rho} = .776$ this gives $\hat{\sigma}^{(N)} = .113$. For large values of n this definition of $\hat{\sigma}^{(N)}$ agrees with (2.2), but in small samples it is more meaningful, being less affected by occasional wild values of the random quantity whose accuracy we are trying to describe.

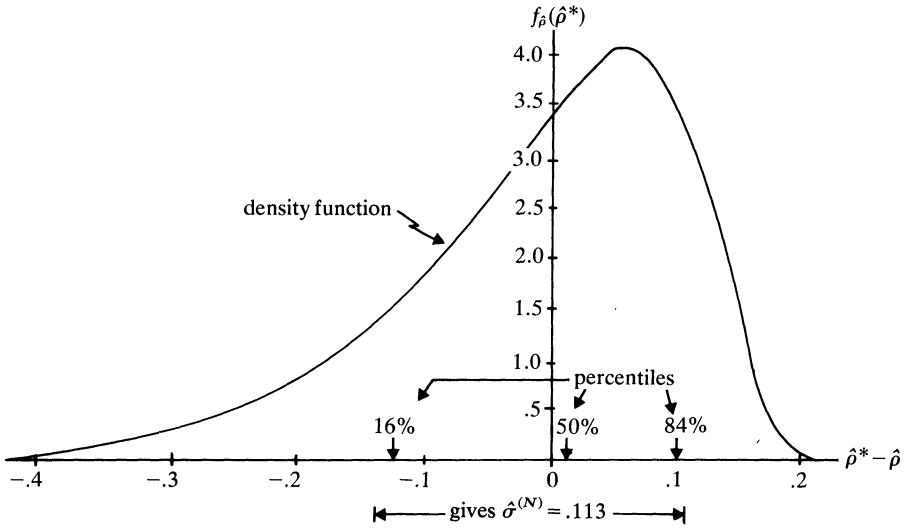

FIG. 2. *The normal theory density function of the observed correlation coefficient* $\hat{\rho}^*$ *for* 15 *data pairs* (x_i, y_i) *drawn from a bivariate normal distribution with true correlation* $\hat{\rho} = .776$. *The distribution puts* 68% *of its probability in the interval* $\hat{\rho}^* \in [\hat{\rho} - .126, \hat{\rho} + .099]$.

The calculations of the next section suggest an answer somewhat closer to $\hat{\sigma}^{(N)} = .113$ than to $\hat{\sigma}^{(J)} = .143$. One bad feature of $\hat{\sigma}^{(J)}$ can be spotted in Table 2. The first value, $\hat{\rho}_{(1)} - \hat{\rho} = .116$, accounts for two-thirds of the sum of squares in (2.3). Any estimate that depends so heavily on a single datum is prone to instability, as we discuss in § 5.

Figure 1 shows why $\hat{\rho}_{(1)} - \hat{\rho}$ is so large. Data point 1 is far away from the other 14, so that its removal causes a large change in the estimated correlation coefficient. This notion is formalized in § 5 under the name "influence function," and furnishes a theoretic rationale for the jackknife estimate of accuracy. In addition to Miller [18], another good reference on the justification and use of the jackknife is Mosteller and Tukey [19].

3. Bootstrap methods. We consider another method, called the "bootstrap" in Efron [8], of assigning an accuracy to the estimated correlation $\hat{\rho} = .776$ for the law school data:[6]

1) Let $\hat{F}$ be the empirical distribution of the 15 observed data points, i.e. the probability distribution which puts mass 1/15 at each observed point (x_i, y_i).

2) Use a random number generator to draw 15 new points (x_i^*, y_i^*) independently and with replacement from $\hat{F}$, so that each new point is an independent random selection of one of the 15 original data points. These new points, which we will call the "bootstrap sample," are a subset of the original points plotted in Fig. 1. Some of the original points will have been selected zero times, some once, some twice, etc.

3) Compute $\hat{\rho}^*$, the correlation coefficient for the bootstrap sample.

4) Repeat steps (2) and (3) a large number of times, say N times, each time using an independent set of new random numbers to generate the new bootstrap sample. Call the resulting sequence of bootstrap correlation coefficients $\hat{\rho}^{*1}, \hat{\rho}^{*2}, \cdots, \hat{\rho}^{*j}, \cdots, \hat{\rho}^{*N}$.

[6] The name "bootstrap" is meant to be euphonic with "jackknife," the two methods being closely related as we shall see, and also to convey the self-help nature of the bootstrap algorithm.

5) Let $[a^*, b^*]$ be the central 68% interval for the $\hat{\rho}^*$ values, i.e.

$$\frac{\#\{\hat{\rho}^{*j} < a^*\}}{N} = .16, \qquad \frac{\#\{\hat{\rho}^{*j} < b^*\}}{N} = .84.$$

Define the bootstrap estimate of the standard deviation σ, say $\hat{\sigma}^{(B)}$, to be half the length of this interval,

$$\hat{\sigma}^{(B)} = \frac{b^* - a^*}{2}.$$

Figure 3 shows the results of $N = 1000$ bootstrap replications. The histogram of the 1000 values $\hat{\rho}^{*1} - \hat{\rho}, \hat{\rho}^{*2} - \hat{\rho}, \cdots, \hat{\rho}^{*N} - \hat{\rho}$, is plotted, and it is seen that $\hat{\sigma}^{(B)} = .127$. The similarity of the histogram to the normal-theory density function of $\hat{\rho}^* - \hat{\rho}$, reproduced from Fig. 2, is apparent, the main difference being an excess of bootstrap values for $\hat{\rho}^* - \hat{\rho} > .15$ (coming from a deficit in the range 0 to .10). This excess pulls the 84% point of $\hat{\rho}^* - \hat{\rho}$ up to .132, compared with the normal-theory value of .099, and is the reason $\hat{\sigma}^{(B)} = .127$ is larger than the normal-theory estimate $\hat{\sigma}^{(N)} = .113$, though it is still considerably smaller than $\hat{\sigma}^{(J)} = .143$, the jackknife estimate.

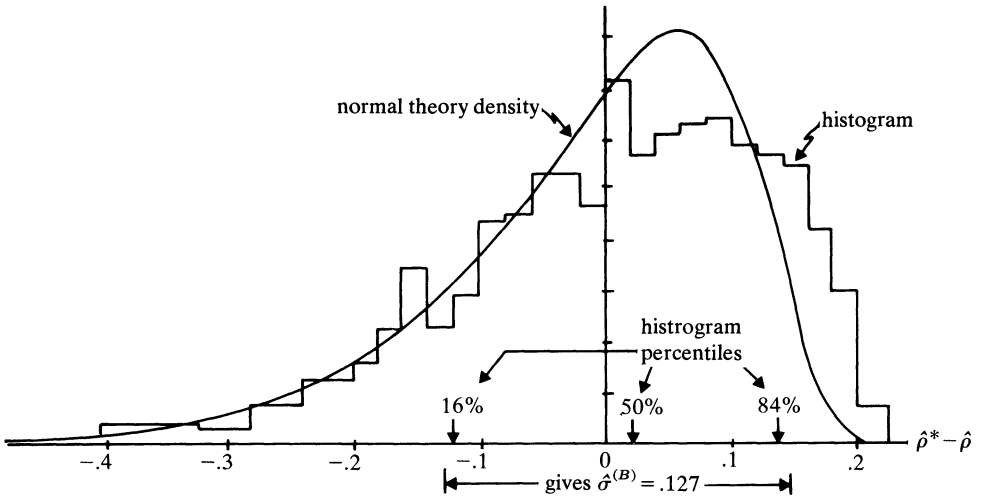

FIG. 3. *Histogram, 1000 bootstrap replications of $\hat{\rho}^* - \hat{\rho}$, gives a bootstrap estimate of accuracy $\hat{\sigma}^{(B)} = .127$ for the correlation coefficient $\hat{\rho} = .776$ of the law school data. The normal theory density of $\hat{\rho}^* - \hat{\rho}$, from Fig. 2, has a similar shape, but falls off more quickly at higher values of $\hat{\rho}^* - \hat{\rho}$.*

What we have called ρ before, the true correlation, might better be called $\rho(F)$, where F is the true probability distribution giving rise to the data pairs (x_i, y_i). The notation $\rho = \rho(F)$ emphasizes that the correlation coefficient is a functional, mapping any bivariate probability distribution into a real number in the interval $[-1, 1]$. Definition (2.1) can be written $\hat{\rho} = \rho(\hat{F})$, where $\hat{F}$ is the empirical probability distribution introduced at step 1 of the bootstrap procedure.

The rationale underlying the bootstrap procedure is simple: i) We want an estimate of the accuracy of $\hat{\rho}$; ii) We would like to use $\sigma(F)$, where $\sigma(\cdot)$ is some agreed upon functional which measures accuracy, such as (2.2), $\sigma(F) = [E(\rho(\hat{F}) - \rho(F))^2]^{1/2}$. (Notice that $\sigma(F)$ depends only upon F since the expectation operator E averages over the possible $\hat{F}$'s arising from a random sample of 15 independent pairs from F.) iii) We don't know F, so instead we estimate $\hat{\sigma} = \sigma(\hat{F})$. In other words, we use the same basic

method to estimate σ as to estimate ρ itself—a simple substitution of $\hat{F}$ for the unknown true distribution F.

Instead of root mean square error, we have been employing a different functional to measure accuracy,

$$\sigma(F) = \text{half the length of the central 68\% of the probability distribution, under } F, \text{ of } \rho(\hat{F}) - \rho(F). \tag{3.1}$$

Why one might prefer (3.1) to (2.2) is discussed in § 5, though the real reason here has been the ease of graphical presentation.

The empirical distribution $\hat{F}$ is a crude estimate of F. Why not use a better estimate of F, say $\hat{F}^+$, and estimate the accuracy by $\hat{\sigma}^+ = \sigma(\hat{F}^+)$? That is exactly what we have done in obtaining the normal theory estimate $\hat{\sigma}^{(N)}$. The better estimate of F is $\hat{F}^+$ equal to a bivariate normal distribution whose correlation coefficient is the observed value $\hat{\rho} = .776$. (The means and variances of $\hat{F}^+$ are also set equal to the observed sample values.) In this sense $\hat{\sigma}^{(N)}$ is itself a bootstrap estimate, the only difference being the use of a better $\hat{F}$ at step 1.[7] "Better," of course, may really be worse if the assumption that the true F is bivariate normal is wrong. It is reassuring to see the agreement between $\hat{\sigma}^{(N)}$ and $\hat{\sigma}^{(B)}$, since the latter makes no special assumptions about the form of F.

It is interesting to try a compromise between $\hat{F}$, the empirical distribution, and $\hat{F}^+$, the best fitting normal distribution. Let $\hat{F}^c$ be the probability distribution of a random point $v = (x, y)$ obtained as follows: take independent points $v' = (x', y')$ and $v'' = (x'', y'')$ from $\hat{F}$ and $\hat{F}^+$ respectively, and let $v = \sqrt{1-c^2}\, v' + cv''$. Then $\hat{F}^0 = \hat{F}$, $\hat{F}^1 = \hat{F}^+$, but for intermediate values of c we get a blend of the discrete distribution $\hat{F}$ and the continuous normal distribution $\hat{F}^+$, which may more nearly approximate our actual beliefs about the form of the true F.

Figure 4 shows what happens if we begin the bootstrap procedure with $\hat{F}^c$ instead of $\hat{F}$. The value $c = 1/\sqrt{5}$ was used, which roughly speaking gives four times as much weight to $\hat{F}$ as to $\hat{F}^+$. The bootstrap distribution of $\hat{\rho}^* - \rho$ now looks even more like the normal theory case, but the estimate of accuracy is virtually unchanged, $\hat{\sigma}^{(B)} = .125$. (An equal mixture of $\hat{F}$ and $\hat{F}^+$ gave $\hat{\sigma}^{(B)} = .116$.) The summary statement $\rho = .776 \pm .125$ seems quite reasonable at this point; we have grounds for believing that the accuracy is somewhat, but not a great deal, worse than the accuracy under pure normal theory.

The choice of $N = 1000$ as the number of bootstrap replications can be shown, in the present case, to determine $\hat{\sigma}^{(B)}$ to an accuracy of about 2.5%. This means that if N were increased from 1000 toward infinity, the limiting value of $\hat{\sigma}^{(B)}$ would be expected to differ from .127 by less than 2.5%. Vastly more bootstrap replications might result in $\hat{\sigma}^{(B)} = .130$ or .125, but almost certainly not $\hat{\sigma}^{(B)} = .120$ or .135. We could have gotten by with $N = 250$ replications, giving an expected accuracy of 5%, but $N = 1000$ is not foolishly excessive. This impressive expenditure of computing power, 1000 times that for the original calculation of $\hat{\rho}$, doesn't include the 1000 smoothed bootstrap replications of Fig. 4. Of course, all the calculations together only took a few seconds and cost perhaps \$10, but, to reiterate the obvious, they would have been practically impossible 30 years ago. Bootstrap-like procedures have undergone very little theoretical development since they have been computationally practical for a

[7] Steps 2 through 5 of the bootstrap procedure are done theoretically, rather than by computer simulation, in the normal-theory calculation. The bivariate normal model is virtually unique in yielding an analytically simple distribution for $\hat{\rho}$. This gets back to our main point, the effect of the computer on what is considered a feasible statistical procedure.

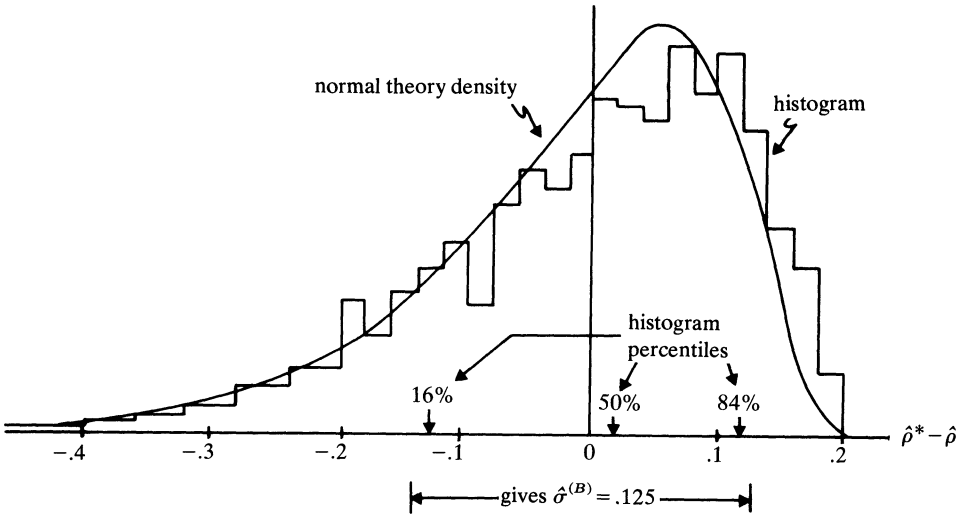

FIG. 4. *Histogram, 1000 bootstrap replications of* $\hat{\rho}^*-\hat{\rho}$, *using the smoothed sampling distribution* $\hat{F}^c$, $c=1/\sqrt{5}$, *described in the text. The histogram follows the normal-theory density more closely than in Fig.* 3, *but* $\hat{\sigma}^{(B)}=.125$, *almost the same value as for the unsmoothed bootstrap.*

comparatively short time, but theoreticians can be expected to take greater interest in them now that they are feasible.

There is an interesting theoretical connection between the jackknife and the bootstrap.[8] Considering now just one bootstrap replication, let p_i^* be the proportion of the bootstrap sample equal to the original data pair (x_i, y_i). For example, if (x_5, y_5) is included three times in the bootstrap sample, of size $n=15$, then $p_5^*=3/15=.20$. The vector $\mathbf{p}^*=(p_1^*, p_2^*, \cdots, p_{15}^*)$ determines $\hat{\rho}^*-\hat{\rho}$, so we can write, say $\hat{\rho}^*-\hat{\rho}=g(\mathbf{p}^*)$, where $g(\cdot)$ is a known function. (To be specific, $g(\mathbf{p}^*)=[\sum p_i^*(x_i-\bar{x}^*)(y-\bar{y}^*)]/[\sum p_i^*(x_i-\bar{x}^*)^2\sum p_i^*(y_i-\bar{y}^*)^2]^{1/2}-\hat{\rho}$, where $\bar{x}^*=\sum p_i^* x_i$, $\bar{y}^*=\sum p_i^* y_i$. Notice that the data (x_i, y_i), $i=1, 2, \cdots, 15$, are considered fixed in this definition.)

The statistics of the vector $\mathbf{p}^*$ are completely known from the properties of the multinomial distribution. For example, $\mathbf{p}^*$ has expected value $(1/n, 1/n, 1/n, \cdots, 1/n)$, $n=15$, and covariance matrix with ijth element

$$\text{Covariance } (p_i^*, p_j^*)=\begin{cases}\dfrac{1}{n^2}-\dfrac{1}{n^3} & i=j\\[2ex] -\dfrac{1}{n^3} & i\neq j.\end{cases} \tag{3.2}$$

Expanding $g(\cdot)$ in a Taylor series around $(1/n, 1/n, \cdots, 1/n)$ gives

$$\hat{\rho}^*-\hat{\rho}=\sum_{i=1}^{n} g^{(i)}\cdot\left(\hat{p}_i^*-\frac{1}{n}\right)+\text{higher order terms}, \tag{3.3}$$

where $g^{(i)}$ is the partial derivative of $g(\cdot)$ with respect to p_i^*, evaluated at $\mathbf{p}^*=(1/n, \cdots, 1/n)$. Together, (3.2) and (3.3) suggest an easy approximation to the standard deviation of $\hat{\rho}^*-\hat{\rho}$ under bootstrap sampling, namely

$$\hat{\sigma}^{(B)}\doteq\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}[g^{(i)}]^2}. \tag{3.4}$$

[8] The remainder of this section assumes some knowledge of statistical theory, though the general drift of the argument still should be discernible to nonstatisticians.

This is almost exactly the jackknife estimate $\hat{\sigma}^{(J)}$, the main difference being the substitution in (2.3) of finite differences[9] in place of the derivatives $g^{(i)}$ appearing in (3.4). Jaeckel [15] originally suggested the right side of (3.4) as an accuracy approximation, calling it the "infinitesimal jackknife;" see also Efron [8].

4. Cross-validation. In its original form, *cross-validation* referred to the following simple, but useful, idea: given a large class of possible models to fit to a set of data, for example linear regression models in which the choice of predictor variables is open to question, first randomly divide the data into two halves. Then fit a model to the first half of the data, using any fitting method at all, and see how well the fitted model predicts the second half of the data. This last step, which is the cross-validation, protects the statistician against an overly optimistic assessment of goodness-of-fit.

Recently many authors, in particular Stone [21] and Geisser [10], have proposed direct use of cross-validation for the selection of appropriate models. This approach is computer intensive, but potentially much broader in application than the familiar linear model approach. We illustrate the method with an example taken from Wahba and Wold [23].

Figure 5 shows 100 artificially generated data points, created according to the following model: the point (x_i, y_i) with abscissa x_i has ordinate y_i randomly determined by

$$y_i = \mu(x_i) + \varepsilon_i \qquad i = 1, 2, \cdots, 100, \tag{4.1}$$

where

$$\mu(x) = 4.26(e^{-x} - 4\,e^{-2x} + e^{-3x}) \tag{4.2}$$

and the ε_i are independent normal random variables with mean 0 and standard deviation $\sigma = 0.2$. The x_i values are equally spaced from 0 to 3.10. The function $\mu(x)$, which in a real application would be unknown to the statistician, is shown as the dashed curve in Fig. 5.

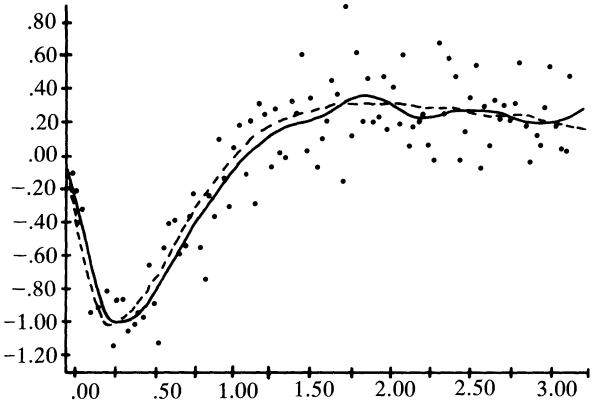

FIG. 5. 100 *random points generated according to model* (4.1), (4.2). *The true mean function* $\mu(x)$ *is indicated by the dashed curve. The solid curve was obtained from the data points by the cross-validation method.*

Wahba and Wold consider fitting a class of curves $\eta(x, \alpha)$ to these data. For a particular choice of the nonnegative parameter α, $\eta(x, \alpha)$ is by definition the curve $\eta(x)$

[9] It is not hard to show that $(n-1)(\hat{\rho} - \hat{\rho}_{(i)})$ approximates $g^{(i)}$; see § 5 of Efron [9].

minimizing

$$\frac{1}{100}\sum_{i=1}^{100}[y_i-\eta(x_i)]^2 \tag{4.3}$$

subject to the constraint

$$\int_0^{3.125}[\eta''(x)]^2\,dx=\alpha. \tag{4.4}$$

Constraint (4.4) is a smoothness condition: if we take $\alpha=0$, $\eta(x,0)$ is very smooth indeed, being the ordinary least squares straight line for the data in Fig. 5. It is easy to see that this gives a very poor fit in the present case. At the opposite extreme, if we let α get large enough, $\eta(x,\alpha)$ will go through every data point. This fits the data perfectly, but is far too irregular a curve to be of any use for prediction or analysis. Intermediate values of α give cubic spline functions, with a trade-off between smoothness (4.4) and fit (4.3).

Cross-validation proposes to estimate the best value of α, without any prior knowledge of the generating mechanism for the data. "Best" here means the value of α minimizing

$$\frac{1}{100}\sum_{i=1}^{100}[\mu(x_i)-\eta(x_i,\alpha)]^2, \tag{4.5}$$

in other words, the curve $\eta(x,\alpha)$ closest to the true mean function $\mu(x)$. Another way to state criterion (4.5) is to imagine that a new set of data, say (x_i, y_i^*), $i=1,2,\cdots,100$, has been independently generated according to model (4.1), (4.2). How well will a curve $\eta(x,\alpha)$ fitted to the original data predict this new data set, in the sense of minimizing $(1/100)\sum_{i=1}^{100}[y_i^*-\eta(x_i,\alpha)]^2$? The expected error of prediction, with $\eta(x,\alpha)$ fixed, is

$$E\frac{1}{100}\sum_{i=1}^{100}[y_i^*-\eta(x_i,\alpha)]^2=\sigma^2+\frac{1}{100}\sum_{i=1}^{100}[\mu(x_i)-\eta(x_i,\alpha)]^2. \tag{4.6}$$

Since $\sigma^2=(0.2)^2$ is a fixed number, minimizing (4.5) is equivalent to minimizing the expected squared error of prediction (4.6).

If the new data set (x_i, y_i^*) were actually available we could easily select α: for each α, the curve $\eta(x,\alpha)$ is determined from the original data set, by (4.3), (4.4), and then tested on the new data set by computing $Q^*(\alpha)=(1/100)\sum_{i=1}^{100}[y_i^*-\eta(x_i,\alpha)]^2$. The α which minimized $Q^*(\alpha)$ would be the estimated best α.

Cross-validation does almost the same thing, without requiring any new data. For each choice of i, $i=1,2,\cdots,100$, let $\eta_{(i)}(x,\alpha)$ be that curve $\eta(x)$ satisfying constraint (4.4), and minimizing

$$\frac{1}{99}\sum_{\substack{j=1\\ j\neq i}}^{100}[y_j-\eta(x_j)]^2. \tag{4.7}$$

In other words, $\eta_{(i)}(x,\alpha)$ is the solution to the constrained minimization problem (4.3), (4.4) with point (x_i, y_i) removed from the data set. We then define

$$Q^{\dagger}(\alpha)=\frac{1}{100}\sum_{i=1}^{100}[y_i-\eta_{(i)}(x_i,\alpha)]^2, \tag{4.8}$$

and select as "best" the α minimizing $Q^{\dagger}(\alpha)$, say $\alpha^{\dagger}$. The curve $\eta(x,\alpha^{\dagger})$ is the proposed estimate for $\mu(x)$.

The solid curve in Fig. 5 shows $\eta(x, \alpha^{\dagger})$ in Wahba and Wold's example. The fit is obviously quite good, and $Q^{\dagger}(\alpha^{\dagger})$, if it were presented, would give a good estimate of the expected prediction error (4.6) for $\eta(x, \alpha^{\dagger})$. Of course we have had to do about 100 times as much work to compute the curve $\eta(x, \alpha)$ for any given α. (Wahba and Wold actually omit points 10 at a time, instead of one at a time, and so reduce the computational effort by a factor of 10.)

Cross-validation resembles the jackknife in that data points are removed one at a time in both procedures, but the underlying connection between the two methods is still not clear to statistical researchers. The next example shows a situation where either cross-validation or the bootstrap can be applied, but the latter is quite a bit more effective. This isn't intended to disparage cross-validation, but rather to suggest that further research may lead to powerful combinations of cross-validation and jackknife–bootstrap methods.

Figure 6 shows 20 artificially generated random points, 10 from each of two populations. The underlying x population is bivariate normal with mean vector $(-\frac{1}{2}, 0)'$

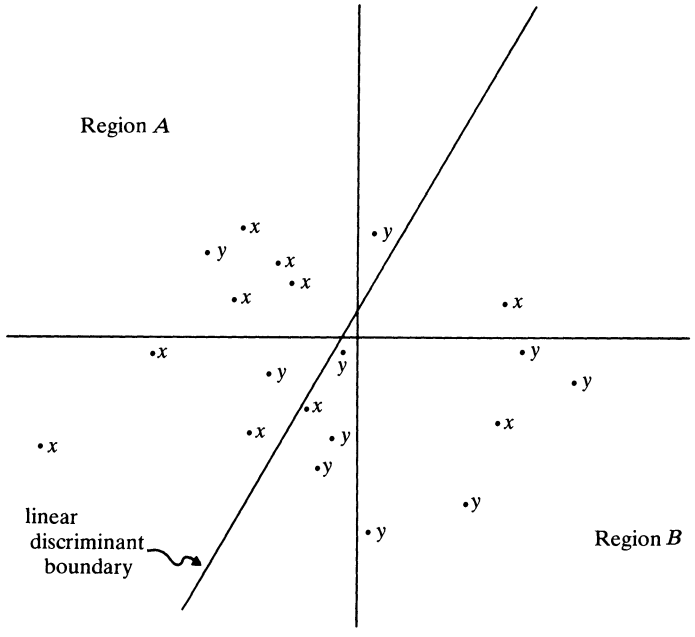

FIG. 6. *Ten x points independently generated from a bivariate normal population with mean vector* $(-\frac{1}{2}, 0)'$, *and ten y points independently generated from a bivariate normal population with mean vector* $(\frac{1}{2}, 0)'$. (*Covariance matrix is the identity in both groups.*) *The straight line is the linear discriminant boundary.*

and covariance matrix the identity. The y population differs in having mean vector $(\frac{1}{2}, 0)'$. By definition, the linear discriminant boundary is the straight line

$$\left\{z: (\bar{y}-\bar{x})'S^{-1}\left(z-\frac{\bar{x}+\bar{y}}{2}\right)=0\right\}, \tag{4.9}$$

where $\bar{x}$ and $\bar{y}$ are the two mean vectors, $\bar{x}=\sum x_i/10$, $\bar{y}=\sum y_i/10$, and S is the 2×2 matrix $\sum (x_i-\bar{x})(x_i-\bar{x})'+\sum (y_i-\bar{y})(y_i-\bar{y})'$. The linear discriminant boundary divides the plane into two regions, A and B, the intention being to classify an unlabeled future point z as being either an x or y depending on whether it falls into A or B. (The optimum division line for future classification is actually $\{z=(z_1, z_2): z_1=0\}$, but of

course the statistician wouldn't know that in a real situation. Notice that the linear discriminant boundary is calculated from the observed data, and doesn't require knowledge of the underlying probability mechanisms. Definition (4.9) is motivated by an attempt to estimate the optimum division line, which is in fact the line obtained from (4.9) when $\bar{x}$, $\bar{y}$ and S are replaced by the true mean vectors and covariance matrix of the two normal populations. Using a linear boundary tacitly assumes that the covariance matrix is the same for both populations.)

The probability that a future x random point will be misclassified is

$$\text{error}_x \equiv \text{Prob}\{x \in B\},$$

which happens to equal 0.41 for the situation in Fig. 6. In this definition, B is considered fixed as shown, and the random quantity is the hypothetical future x point. The obvious estimate of error_x is

$$\widehat{\text{error}}_x = \frac{\#\{x_i \in B\}}{10},$$

which equals 0.30 in Fig. 6. It is well known that $\widehat{\text{error}}_x$ tends to underestimate error_x, that is to have an optimistic bias, and an important problem is to estimate the expected bias,

$$\text{bias}_x \equiv E\{\text{error}_x - \widehat{\text{error}}_x\}. \tag{4.10}$$

The corresponding quantity for the y population is equally important of course, but it is sufficient to discuss estimating bias_x.

Cross-validation estimates bias_x by i) successively eliminating each point x_i, $i = 1, 2, \cdots, 10$; ii) recomputing the linear discriminant boundary on the basis of the nine remaining x's and 10 y's; and iii) seeing whether or not x_i is misclassified by the recomputed discrimination rule. Let $\text{error}_x^{\dagger}$ be the proportion of the x points misclassified at step iii). Then the cross-validated estimate of bias is

$$\text{bias}_x^{\dagger} = \text{error}_x^{\dagger} - \widehat{\text{error}}_x. \tag{4.11}$$

In the situation of Fig. 6, $\text{bias}_x^{\dagger} = 0.10$, which means that 4 out of 10 x values were misclassified during the cross-validation process.

The bootstrap estimate of bias_x takes considerably more computation:

1) Select a bootstrap sample of 10 new x points, $x_1^*, x_2^*, \cdots, x_{10}^*$, by random sampling, independently and with replacement, from the given points $x_1, x_2, \cdots, x_{10}$. Likewise, construct a bootstrap sample of 10 new y points $y_1^*, y_2^*, \cdots, y_{10}^*$ by random sampling from $y_1, y_2, \cdots, y_{10}$.

2) Construct the bootstrap linear discriminant boundary by substituting $\bar{x}^*$, $\bar{y}^*$, S^* for $\bar{x}$, $\bar{y}$, S in (4.9). Denote the bootstrap discriminant regions as A^*, B^*.

3) Let

$$b_x^* \equiv \frac{\#\{x_i \in B^*\}}{10} - \frac{\#\{x_i^* \in B^*\}}{10}. \tag{4.12}$$

4) Repeat steps 1)–3) a large number N of times, obtaining independent values $b_x^{*1}, b_x^{*2}, \cdots, b_x^{*N}$, and estimate bias_x by

$$\text{bias}_x^* = \frac{1}{N}\sum_{j=1}^{N} b_x^{*j}. \tag{4.13}$$

In the present case, $N = 100$ bootstrap replications gave the estimate $\text{bias}_x^* = 0.078$. Notice that (4.12) is of the form "true minus apparent error rate," where now "true" refers to the x_i and "apparent" refers to the x_i^*. The justification of the bootstrap is the same here as in § 3.

When the x and y values are generated by the underlying normal distributions described earlier, the actual value of bias_x is .062. That is, $\widehat{\text{error}}_x$ tends to underestimate error_x by .062, on the average. In a large number of Monte Carlo trials, reported in § 4 of Efron [8], both $\text{bias}_x^\dagger$ and bias_x^* were themselves nearly unbiased; that is they averaged about .062. However, the $\text{bias}_x^\dagger$ values were three times more variable than the bias_x^* values, which made them much less dependable for assessing bias_x in any particular case.

5. Robust estimation. A fundamental statistical tactic is the combination of separate small pieces of information, each by itself nearly worthless, to produce an overall conclusion of substantial reliability. Independent tosses of a possibly biased coin offer the classic example. No one toss tells us very much about the coin, but having observed, say, 30 heads in 100 tosses, the true probability of heads can reliably be predicted to lie in the interval $.300 \pm .092$. Averaging, which is what is done to get the estimate .300, is a powerful way of bringing diverse information to bear on a single important question. Some of the most useful statistical methods, such as linear regression and analysis of variance, are really no more than fancy averaging techniques, designed for situations where the individual observations are collected under varying circumstances.

Suppose we threw away any one of the 100 coin flips, leaving ourselves with the data from the remaining 99. The estimated true probability of heads, call it p, would then equal either $p = 30/99 = .303$ or $p = 29/99 = .293$, depending on whether we had thrown away a head or a tail. Both .303 and .293 are quite close to .300, the point here being that no one of the individual pieces of information is by itself very important to the estimate $p = .300$. We say that p is *robust* in this situation, to use Tukey's memorable terminology (somewhat differently than originally intended).

Unfortunately, it is not always true that the average $\bar{x} = \sum_{i=1}^n x_i/n$ is robust in the sense above. Table 3 shows microbe counts in 69 swabs from different portions of a Mariner space probe. The average count is $\bar{x} = 16.14$, but deleting the largest count, count $\#69$, gives average $\bar{x}_{(69)} = 1.53$ for the remaining 68 numbers. Deleting the largest two counts, count $\#69$ and count $\#68$, gives $\bar{x}_{(68,69)} = .63$. In this case $\bar{x}$ is distinctly nonrobust.

Recently statisticians have become interested in robust estimators, averaging techniques which limit the influence of any one observation on the estimate, even in situations as extreme as that of Table 3. Huber's monograph [14] gives an excellent overview of the subject. Another good reference is Hampel [12].

TABLE 3

Microbe counts in 69 *swabs of a Mariner space probe. (Part of a much larger data set.) The count was zero in* 53 *swabs, one in* 6 *swabs, etc. Removing the largest count,* 1010, *reduces the average count from* 16.14 *to* 1.53.

Count	0	1	3	4	5	6	9	62	1010
Number of swabs	53	6	4	1	1	1	1	1	1

The average $\bar{x}$ of a set of numbers $x_1, x_2, \cdots, x_n$ can also be derived as that number T which minimizes the sum of squared deviations, $\sum_{i=1}^{n} (x_i - T)^2$. Differentiation shows that $\bar{x}$ may also be characterized as the solution to the equation (in T), $\sum_{i=1}^{n} (x_i - T) = 0$. By definition, an *M estimator* is the solution in T to the equation

$$\sum_{i=1}^{n} \psi(x_i - T) = 0. \tag{5.1}$$

Here $\psi(\cdot)$ is a preselected function, which can be chosen to give good robustness properties. If $\psi(x) \equiv x$ then T is the ordinary average $\bar{x}$. If

$$\psi(x) = \text{sign } (x)$$

then T is the sample median, the middle value of the observations listed in increasing order. (Reversing the differentiation argument at the beginning of this paragraph shows that the median minimizes the sum of absolute deviations $\sum_{i=1}^{n} |x_i - T|$.) For the microbe data the median equals 0 no matter how many of the nonzero counts are removed. This is more robustness than we want in many situations!

As a compromise between $\psi(x) = x$ and $\psi(x) = \text{sign } (x)$ we can take

$$\psi(x) = \begin{cases} -c, & x < -c, \\ x, & -c \leqq x \leqq c, \\ c, & c < x. \end{cases} \tag{5.2}$$

Choosing $c = \infty$ makes T equal to the average, while $c = 0$ (actually, the limit as $c \to 0$, in which case $\psi(x)/c \to \text{sign } (x)$), gives the median. The choice $c = 10$ results in the estimate $T = .93$ for the microbe data. Removing the largest count changes the estimate to $T_{(69)} = .78$; also removing the second largest gives $T_{(68,69)} = .63$. These values can be obtained easily on a hand calculator, using Newton–Raphson iteration or just trial and error. Doing the computation gives a good feeling for the way in which the estimator based on (5.2) acts like $\bar{x}$ near the middle of the data, but automatically limits the influence of outlying observations.

How can we choose amongst possible estimators T in any given situation? If we knew that the observations were independently generated according to some probability density function $f(x - \theta)$, with θ an unknown parameter to be estimated (a "translation family" situation), we could use the maximum likelihood estimator, i.e. the number T which maximizes $\prod_{i=1}^{n} f(x_i - T)$. Taking logarithms and differentiating shows that the maximum likelihood estimator is an M estimator,[10] with ψ function equal to

$$\psi_f(x) \equiv -\frac{f'(x)}{f(x)}. \tag{5.3}$$

For the normal translation family, with $f(x - \theta) = (2\pi)^{-1/2} \exp\{-\frac{1}{2}(x - \theta)^2\}$, $\psi_f(x) = x$ and so the average $\bar{x}$ is the maximum likelihood estimator. The Laplace translation family $f(x - \theta) = (\frac{1}{2}) \exp\{-|x - \theta|\}$ gives the median as the maximum likelihood estimator. Maximum likelihood produces nearly optimal estimates in translation families, assuming of course that the $f(\cdot)$ used in (5.3) is actually the correct form of the density function.

The point of much of the work in robustness theory is that the statistician may not completely trust a given parametric model, such as the normal translation family, and so

[10] The name "M estimator" comes from Maximum likelihood.

may prefer to change the "optimum" $\psi_f(\cdot)$ to a more robust choice $\psi(\cdot)$. This reduces the theoretical efficiency of the estimator somewhat, compared with that of the maximum likelihood estimator, if the model $f(x-\theta)$ is correct, but protects the statistician against disastrously foolish estimates if the model is somewhat off. It can be shown that the square of the correlation between $\psi_f(x)$ and $\psi(x)$ calculated under density $f(x)$, determines the large-sample efficiency of the M estimator based on $\psi(\cdot)$. A correlation of .90, for example, means that the M estimator based on $\psi(\cdot)$ wastes about 19% $(.19 = 1-.9^2)$ of the information available for estimating θ under the model $f(x-\theta)$. It turns out that for the normal translation model, reasonable choices of c in (5.2) give efficiencies better than 95% while still providing good protection against occasional wild observations.

We have discussed the *influence* of a single observation on an estimate T. This notion has been formalized under the name "influence function," and provides theoretical justification for the jackknife, as well as for robust estimators. The M estimators are functionals $T(\hat{F})$, as was $\rho(\hat{F})$ in § 3, and can be thought of as estimating the true value $T(F)$, where F is the true probability distribution giving rise to the data $x_1, x_2, x_3, \cdots$. If the sample size n were increased toward infinity, $T(\hat{F})$ would approach $T(F)$.

Let δ_x represent the degenerate probability distribution putting all of its mass at the point x. The influence function $\dot{T}(x;F)$, for a given estimator T, evaluated at the true distribution F, is the function of x defined by

$$\dot{T}(x;F) \equiv \frac{d}{d\varepsilon} T((1-\varepsilon)F + \varepsilon\delta_x)\Big|_{\varepsilon=0} \tag{5.4}$$

The influence function represents the effect upon $T(F)$ of a small local change in F. By superimposing many such small changes we obtain, via a first order Taylor series expansion, an approximation to $T(\hat{F}) - T(F)$, the difference between the estimated and true value of T,

$$T(\hat{F}) \doteq T(F) + \frac{1}{n}\sum_{i=1}^{n} \dot{T}(x_i;F). \tag{5.5}$$

For a linear functional, such as the mean $T(F) = \int x\, dF(x)$, (5.5) is exact. (For the mean, $\dot{T}(x;F) = x - T(F)$, so $(1/n)\sum_{i=1}^{n} \dot{T}(x_i;F) = \bar{x} - T(F) = T(\hat{F}) - T(F)$.) Nonlinear functionals, such as the M estimator based on (5.2), are, under some regularity conditions, asymptotically linear, as $n \to \infty$, in the sense of (5.5). The usefulness of (5.5) is that it approximates $T(\hat{F}) - T(F)$ by the average of independent, identically distributed, random quantities $\dot{T}(x_i;F)$. The standard deviation of such an average is

$$\frac{[\int [\dot{T}(x;F)]^2\, dF(x)]^{1/2}}{\sqrt{n}}, \tag{5.6}$$

$1/\sqrt{n}$ times the root mean square of the influence function. The jackknife standard deviation $\hat{\sigma}^{(J)}$, (2.3), is the nonparametric estimate of (5.6). (The values $\hat{\rho}_{(i)} - \hat{\rho}$ are rather crude estimates of the influence function. Expression (5.6) is closely related to (3.4).)

The principle of robust estimation can now be stated more quantitatively: only use estimators $T(\hat{F})$ for which the influence function is sensibly bounded. It is easy to verify that the influence function of an M estimator is proportional to $\psi(x - T(F))$. The form of ψ in (5.2) is nothing more than a modification of ψ for the average, $\psi(x) = x$, with a

bound put on the magnitude of the influence function. Definition (3.1) is motivated by similar considerations.

Robustness ideas are now being applied to regression situations such as (4.1), nice references being Andrews [1] and Mosteller and Tukey [19]. If the errors ε_i occasionally take on wild values, then fitting models by the method of least squares can go disastrously wrong. The least squares method fits regression parameters β (for example, the coefficients of the exponential terms in (4.2), if they were unknown) by minimizing $\sum_{i=1}^{n}(y_i-\mu_\beta(x_i))^2$. Instead, we can minimize $\sum_{i=1}^{n}\Psi(y_i-\mu_\beta(x_i))$, where

$$\Psi(y)=\int_0^y \psi(y')\,dy', \tag{5.7}$$

with ψ as in (5.2). The limiting case, as $c\to 0$, fits a model by minimizing $\sum_{i=1}^{n}|y_i-\mu_\beta(x_i)|$, the sum of absolute deviations. "Least absolute deviations" was the fitting method favored by Laplace, but it lost out to Gauss' least squares, mainly on the grounds of computational simplicity. Now, 150 years later, Laplace may reclaim the field, with the assistance of the modern computer.

6. Censored data. We have made frequent use of the empirical distribution $\hat{F}$, the probability distribution which puts mass $1/n$ at each of n observed data points $x_1, x_2, \cdots, x_n$. (In §§ 3 and 4 the x_i were points in a two dimensional space, while in § 5 the space was one dimensional.) It may seem that there is no way to make the calculation of $\hat{F}$ difficult. If so, a look at some *censored data* should convince the reader otherwise.

Table 4 shows some early results from the heart transplant program at Stanford. The survival times in days, following the transplant operation, are listed for 18 patients. The first listed patient survived 3 days, the second 4+ days, where the "+" indicates that the patient was still alive on April 13, 1972, the point in time at which the data was collected. Here it would be wrong to let $\hat{F}$ be the distribution putting mass $1/18$ at each of the numbers 3, 4, 10, 25, $\cdots$, 1025 since, for eample, the actual survival time corresponding to 4+ is known only to lie in the interval $(4,\infty)$. This is an example of *censoring*, in which the exact value of a measurement can't be seen, but some information on its whereabouts is available.

Let T represent the survival time of a heart transplant patient, a quantity which we will measure in days. The *survival curve* $S(t)$ is the probability of surviving past a given time t,

$$S(t)\equiv \text{Prob}\{T>t\}. \tag{6.1}$$

Knowing the function $S(t)$ is the same as knowing F, the true probability distribution of T. If there were no censoring we could construct an estimate of $S(t)$ in the obvious way,

$$\hat{S}(t)=\frac{\#\{T_i>t\}}{n}, \tag{6.2}$$

where $n=18$ in the case above. In other words, we could use the ordinary estimate $\hat{F}$, of which $\hat{S}(t)$ is another representation.

Figure 7 shows how $\hat{S}(t)$ is constructed when some of the data are censored. The construction depends upon the number of patients at *risk* at time t,

$$n(t)\equiv \text{number of patients neither censored nor observed to die before time } t, \tag{6.3}$$

which is given in Table 4. In our example, $n(0)=18$, $n(100)=9$, $n(200)=4$, etc. The

TABLE 4

Survival times for 18 early heart transplant patients. Tabled is survival time, in days, following the transplant; "+" indicates that the patient was still alive on April 13, 1972, the day the data were collected. Abstracted from a larger data set in Brown and Turnbull [21]. "Number at risk" is used in the calculation of $\hat{F}$.

Survival time	3	4+	10	25+	39	40+	43	54	65
Number at risk	18	17	16	15	14	13	12	11	10
Survival time	120+	136	147	157+	183+	312	546+	824	1025
Number at risk	9	8	7	6	5	4	3	2	1

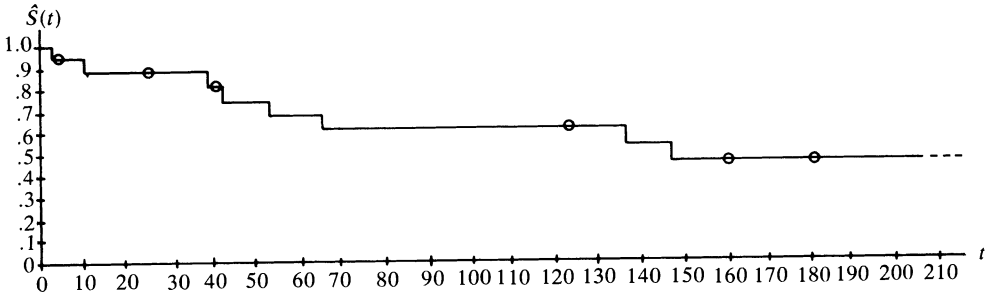

FIG. 7. *Estimated survival curve $\hat{S}(t)$ from the data in Table 4. Open circles represent censored data points, while jumps occur at uncensored observations. At each uncensored data point, i.e. at each observed death, $\hat{S}(t)$ is multiplied by a factor equal to the proportion of the observable population not dying.*

definition of $\hat{S}(t)$ is recursive, starting with $\hat{S}(0) = S(0) = 1$:

$$\hat{S}(t) = \begin{cases} \hat{S}(t-1) & \text{if no observed deaths on day } t \\ \hat{S}(t-1)\dfrac{n(t)-1}{n(t)} & \text{if one observed death on day } t. \end{cases} \tag{6.4}$$

In the heart transplant example $\hat{S}(2) = \hat{S}(1) = \hat{S}(0) = 1$, $\hat{S}(3) = \hat{S}(2)(17/18) = .944$, $\hat{S}(10) = \hat{S}(9)(15/16) = .994 \cdot .938 = .885$, etc. Notice that the data point 4+ figures in the denominator of $\hat{S}(3)$, but has no further effect on $\hat{S}(t)$. Kaplan and Meier [16] give a very readable account of the theory behind (6.4).

The construction of $\hat{S}(t)$ may seem ad hoc, but Kaplan and Meier show that it produces the maximum likelihood estimate of the unknown $S(t)$: among all possible survival curves $S(t)$, i.e. among all possible true distributions F, the choice $S(t) = \hat{S}(t)$ maximizes the probability of obtaining the data actually observed. Bootstrap estimates of accuracy for functionals of censored data begin with $\hat{F}$ corresponding to the survival curve $\hat{S}(t)$, at step 1 of the bootstrap algorithm.

Efron [7] suggested another motivation for $\hat{S}(t)$. Suppose we start out with any estimate $\hat{S}^{(0)}(t)$. Define a new estimate $\hat{S}^{(1)}(t)$, along the lines of (6.2),

$$\hat{S}^{(1)}(t) = \frac{\hat{E}^{(0)}(\#\{T_i > t\})}{n}, \tag{6.5}$$

where $\hat{E}^{(0)}$ indicates an expectation taken with respect to the probability distribution defined by the survival curve $\hat{S}^{(0)}(t)$. Taking $t=20$ in Table 4, for example, the 15 patients with survival times >20, censored or not, contribute 15 to the $\#\{T_i>20\}$. The patients with survival times 3 and 10 contribute zero to $\#\{T_i>20\}$. The patient with survival time 4+ may or may not have $T_i>20$. This patient contributes an expected amount to the right side of (6.5), the expectation being taken under the distribution $\hat{S}^{(0)}(t)$, so that $\hat{S}^{(1)}(20)$ is between 15/18 and 16/18.

We can iterate (6.5), giving the sequence of survival curves $\hat{S}^{(0)}(t)$, $\hat{S}^{(1)}(t)$, $\hat{S}^{(2)}(t), \cdots$. Efron shows that this sequence converges to $\hat{S}(t)$. The usefulness of this iterative construction of $\hat{S}(t)$ is that it can be applied under more difficult censoring conditions. The data in Table 4 consists of observed deaths and *right-censored* observations, such as 4+. In other situations there may also be left-censored and doubly censored observations ("the event occurred before $t=17$," "the event did not occur during the interval (12, 20)"). Turnbull [22] showed that under general censoring conditions, the iterative construction (6.5) always converges to the maximum likelihood estimate of $S(t)$.

Recently, Dempster, Laird, and Rubin [6] have put the relationship between (6.5) and maximum likelihood estimation into a wider context, unifying work by many earlier writers. They consider a variety of situations in which it would be easy to calculate the maximum likelihood estimator if one had the full set of data, but where for one reason or another some of the information is missing. An example, for those familiar with analysis of variance, is a two way table with a few missing observations. In such a situation they show that an iterative procedure like (6.5) always leads to the maximum likelihood estimator, and moreover does so in a monotonic manner. They call this method the "EM Algorithm," in which one first Estimates the missing data and then Maximizes as if the full data set were present.

The survival function $S(t)$ can be expressed as

$$S(t)=\prod_{s=1}^{t}[1-h(s)], \tag{6.6}$$

where

$$h(s)\equiv \text{Prob}\{T=s \mid T>s-1\}, \tag{6.7}$$

the conditional probability of dying on day s given survival past day $s-1$. The function $h(s)$ is called the *hazard rate*. The estimate (6.4) comes from estimating the factor $1-h(s)$ by

$$1-\frac{\text{number of observed deaths on day } s}{n(s)}. \tag{6.8}$$

Hazard rates are more convenient to work with than density functions in censored data situations, an idea we now explore further.

We have treated the patients in Table 4 as if they were identical, at least as far as the probability distribution of their survival times is concerned. In fact, there are observable differences between the patients—age, sex, race, etc.—which we might wish to examine for their effect on survival time. If there were no censoring we could run an ordinary regression analysis with the observed survival times as the dependent variable. Cox [4] has suggested a regression analysis which works directly with the hazard rates, and is not affected by data censoring.

Let z_i represent the vector of relevant observable information, such as age, race, and sex, about patient i, coded in some fashion so that all the entries of z_i are numbers.

For example, a 57 year old white male might be coded (57, 0, 1) where "0" indicates white and "1" indicates male. Cox's model postulates that the hazard rate for patient i, say $h_i(s)$, is of the form

$$h_i(s) = g(s)\, e^{\beta' z_i}. \tag{6.9}$$

Here $g(s)$ is an overall hazard rate applying to all the patients and β is a vector of unknown coefficients, corresponding to the regression coefficients in an ordinary regression model. If $\beta = 0$ then all the patients have the same hazard rate, i.e. identical probability distributions for their survival times, but if $\beta \neq 0$, model (6.9) says that the survival time distributions are functions of z_i. (The vector z_i can itself be a time varying function, say $z_i(s)$, as long as it is always observable.)

In order to analyze this model, Cox uses an approach similar to (6.8). Let $\mathcal{R}(s)$ be the *risk set* on day s, the set of patients available for observation on that day, i.e. those who have not been previously censored nor observed to die. Given that there is one death on day s, the probability under model (6.9) that it was some particular patient in $\mathcal{R}(s)$ who dies, say patient i_s, equals

$$\frac{e^{\beta' z_{i_s}}}{\sum_{i \in \mathcal{R}(s)} e^{\beta' z_i}}. \tag{6.10}$$

(Expression (6.10) is actually an approximation which becomes exact as the units in which we are measuring time become infinitesimal.)

The advantage of (6.10) is that it depends only on β and the observable vectors z_i, and not on the common hazard function $g(s)$ in (6.9). This makes it easy to analyze the data for the effects of β, without any modeling of $g(s)$ being necessary. Cox multiplies the factors (6.10) together, one from each observed death,

$$\prod_{\substack{\text{observed} \\ \text{deaths}}} \frac{e^{\beta' z_{i_s}}}{\sum_{i \in \mathcal{R}(s)} e^{\beta' z_i}}, \tag{6.11}$$

and treats this product as if it were an ordinary likelihood function for β. For example, the β which maximizes (6.11) is treated as a maximum likelihood estimate. This approach ignores part of the data, those days on which no deaths occur, but has been shown to give reasonably efficient estimates of β nevertheless.

If there are many patients, a hundred or more, and the vectors z_i are time-varying, expression (6.11) can be computationally quite difficult to deal with, taxing even a large computer. Without a computer, the method is hopeless, except in the simplest situations. (Mantel and Haenszel [17] discuss one such situation, the two sample comparison problem of § 1, with censored data.) Cox's regression method is a good example of a statistical theory which has developed in response to the capacity of modern computational equipment.

7. Conclusion. The purpose of mathematical theory, and in fact all scientific theory, is to reduce complicated situations to simple ones. Just what a scientist means by "simple" is determined by experience, training, convention, and the limitations of human reasoning faculties. A Taylor series expansion is a classic example of this process: a given function is expressed as a sum of multiples of powers. Since we are taught a lot about sums, multiples, and powers, the explanation may be a good deal easier to understand than the function as originally stated.

The advent of the high speed computer has redefined "simple" in the mathematical sciences. For example, an optimization problem which can be reduced to a problem in

linear programming is, in most instances, now considered solved, since the simplex method is so efficient in numerically solving linear programs.

The purpose of this article has been to show this same process at work in mathematical statistics. A theory which enables a scientist to understand his data with the help of a high speed computer may now be as useful as a theory which only requires a table of the exponential function, particularly if the latter theory does not exist. Computer assisted theory is no less "mathematical" than the theory of the past, it is just less constrained by the limitations of the human brain.

The need for a more flexible, realistic, and dependable statistical theory is pressing, given the mountains of data now being amassed. The prospect for success is bright, but I believe the solution is likely to lie along the lines suggested in the previous sections—a blend of traditional mathematical thinking combined with the numerical and organizational aptitude of the computer.

REFERENCES

[1] D. F. ANDREWS, *A robust method for multiple linear regression*, Technometrics 16 (1974), pp. 523–531.

[2] B. W. BROWN AND B. W. Turnbull, *Survivorship analysis of heart transplant data*, Department of Statistics, Stanford University, Technical Report No. 34 (1972).

[3] B. W. BROWN, B. W. TURNBULL AND M. HU, *Survivorship analysis of heart transplant data*, Journal of the American Statistical Association, 69 (1974), pp. 74–80.

[4] D. R. COX, *Regression models and life-tables*, Journal of the Royal Statistical Society Series B, 34 (1972), pp. 187–220.

[5] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1946.

[6] A. P. DEMPSTER, N. M. LAIRD AND D. B. RUBIN, *Maximum likelihood estimation from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society Series B 39 (1977), pp. 1–38.

[7] B. EFRON, *The two sample problem with censored data*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability IV, (1967), pp. 831–853.

[8] ———, *Bootstrap methods: Another look at the jackknife*, Annals of Statistics, 7 (1979), no. 1, to appear.

[9] ———, *Controversies in the foundations of statistics*, American Mathematical Monthly 85 (1978), no. 4, pp. 231–246.

[10] S. GEISSER, *The predictive sample reuse method with applications*, Journal of the American Statistical Association 70 (1975), pp. 320–328.

[11] G. H. GOLUB AND G. P. H. STYAN, *Numerical computations for univariate linear models*, Journal of Statistical Computation and Simulation 2 (1973), pp. 253–274.

[12] F. R. HAMPEL, *The influence curve and its role in robust estimation*, Journal of the American Statistical Association 69 (1974), pp. 383–393.

[13] P. J. HUBER, *Robust statistics: a review*, Annals of Mathematical Statistics 43 (1972), pp. 1041–1067.

[14] ———, *Robust Statistical Procedures*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1977.

[15] L. JAECKEL, *The infinitesimal jackknife*, Bell Labs. Memorandum #MM 72-1215-11, 1972.

[16] E. L. KAPLAN AND P. MEIER, *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association 53 (1958), pp. 457–481.

[17] N. MANTEL AND W. HAENSZEL, *Statistical aspects of the analysis of data from retrospective studies of disease*, Journal of the National Cancer Institute 22 (1959), pp. 719–748.

[18] R. G. MILLER, *The jackknife—a review*, Biometrika 61 (1974), pp. 1–17.

[19] F. MOSTELLER AND J. W. TUKEY, *Data Analysis and Regression*, Addison-Wesley, Reading, MA, 1977.

[20] D. B. RUBIN, *Using empirical Bayes techniques in the law school validity studies*, Law School Admission Council Report 78-1, 1977.

[21] M. STONE, *Cross-validatory choice and assessment of statistical predictions*, Journal of the Royal Statistical Society Series B 36 (1974), pp. 111–147.

[22] B. W. TURNBULL, *The empirical distribution function with arbitrarily grouped, censored, and truncated data*, Journal of the Royal Statistical Society Series B 38 (1976), pp. 290–295.

[23] G. WAHBA AND S. WOLD, *A completely automatic French curve: fitting spline functions by cross-validation*, Communications in Statistics 4 (1975), pp. 1–17.