The Annals of Statistics
1982, Vol. 10, No. 2, 340–356

THE 1981 WALD MEMORIAL LECTURES

MAXIMUM LIKELIHOOD AND DECISION THEORY

BY BRADLEY EFRON

Stanford University

This paper discusses five questions concerning maximum likelihood estimation: What kind of theory is maximum likelihood? How is maximum likelihood used in practice? To what extent can this theory and practice be justified from a decision-theoretic viewpoint? What are maximum likelihood's principal virtues and defects? What improvements have been suggested by decision theory?

1. Purpose. Maximum likelihood is the most widely used statistical estimation technique. It emerged in modern form 60 years ago in a series of remarkable papers by Fisher (1922, 1925, 1934). The surprising fact is that maximum likelihood is still the source of considerable controversy in the statistical community, as seen in Berkson's 1980 paper "Minimum chi-square, not maximum likelihood!" and the ensuing discussion.

The controversy centers on the relationship between decision theory and maximum likelihood. Beginning with the Neyman-Pearson lemma, decision theory has reshaped the theory and practice of hypothesis testing. The same cannot be said of estimation. Maximum likelihood continues to dominate statistical practice, essentially in its original formulation, not much affected by Waldian developments.

This paper concerns five main questions. (1) If maximum likelihood isn't decision theory, then what kind of theory is it? (2) How is maximum likelihood used in practice? (3) To what extent can this theory and practice be justified from a decision-theoretic viewpoint? (4) What are maximum likelihood's principal virtues and defects? (5) What genuine improvements have been suggested by decision theory?

Technical details are kept to a minimum in what follows. None of the technical points are new, many of them dating back to Fisher. A good reference is Cox and Hinkley (1974, Chapter 9). The discussion here is written entirely from a frequentist viewpoint. Every attempt has been made to avoid the usual Bayesian-frequentist-Fisherian arguments. The basic issue is more practical than philosophical: what does a statistician do when faced with a new body of data? Maximum likelihood provides a practical way to begin and carry out an analysis. We discuss the advantages and drawbacks of this program.

2. An example. Figure 1 compares the field goal kicking ability of Don Cockroft, kicker for the Cleveland Browns, with the aggregated data for all field goal kickers in the American Football Conference (AFC), 1969–1972. The data, which are taken from published records of the National Football League, lists attempts and successes from various distances. For instance, Cockroft successfully completed 15 of 32 attempts from between 30 and 39 yards out, compared to 238 of 372 for the entire AFC. Over all yardages Cockroft completed 56 out of 100 attempts, compared to 891 successes out of 1494 attempts for the AFC.

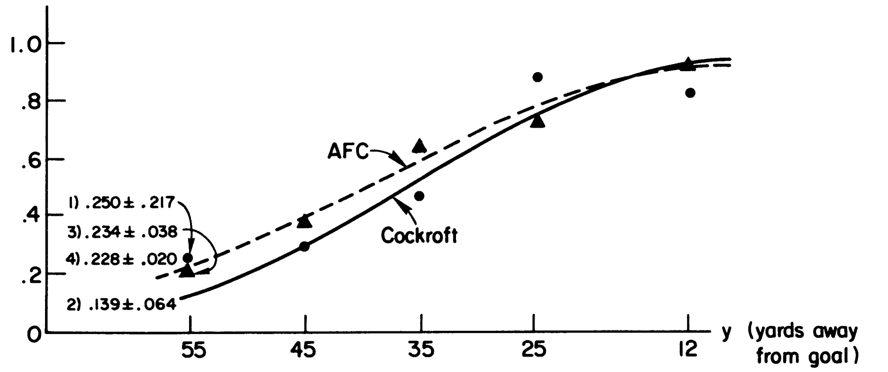
Also appearing in Figure 1 are estimated logistic regressions for the probability of

Received September 1981; revised November 1981.

Key words and phrases. Maximum likelihood estimate, summary, Fisher information, Stein estimation.

AMS 1980 *subject classifications.* 62A10, 62B10.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org



Cockroft	●	1/4	8/27	15/32	22/25	10/12	56/100
AFC	▲	29/124	176/449	238/372	222/306	226/243	891/1494

FIG. 1. *Field goal attempts and successes,* 1969–1972, *at various distances (yards away) from goal. Circles indicate Cockroft's success proportions, triangles proportions for the entire* AFC. *Solid curve is the fitted logistic regression for Cockroft, dashed curve is the logistic regression for the* AFC.

success. Letting γ_y be the probability of success from y yards out, the fitted model is

$$\gamma_y = \frac{1}{1 + \exp\{-\alpha - \beta(y - 30)\}}. \tag{2.1}$$

Since the published data is grouped, the following code for y was used in fitting (2.1) and will be assumed in what follows:

0–19 yards, $y = 12$;	40–49 yards, $y = 45$;
20–29 yards, $y = 25$;	50+ yards, $y = 55$.
30–39 yards, $y = 35$.	

The solid curve is the regression (2.1) fitted by maximum likelihood to Cockroft's 100 data points, the dashed curve is the corresponding regression fitted to all 1494 data points.

To focus discussion, we consider the following deliberately simplified problem: estimate

$$\gamma_{55} = \text{Cockroft's probability of success from 55 yards out.} \tag{2.2}$$

Four different maximum likelihood estimates are indicated in Figure 1. (1) The estimate based on Cockroft's four attempts from 55 yards, .250 ± .217. (2) The estimate based on the logistic regression fit to all 100 Cockroft attempts, evaluated at $y = 55$, .139 ± .064. (3) The estimate based on the AFC's 124 attempts from 55 yards, .234 ± .038. (4) The estimate based on the logistic regression fit to all 1494 AFC attempts, .228 ± .020. These estimates are all obtained by applying standard maximum likelihood theory, as described in Section 4.

3. What is "estimation"? In order to discuss maximum likelihood, we have to describe what "estimation" means in a typical statistical situation such as the field goal example. (See Barnard (1974) for an interesting dicussion of the same question.) Figure 2 schematically illustrates the major steps in a data analysis, and the place of estimation theory.

The most basic statistical process is *enumeration*, the collecting and listing of individual cases. The fundamental idea of statistics is that useful information can be accrued from individual small bits of data. No one field goal try by itself tells us much, but together the data speak clearly. A statistician looking at Figure 1 knows more about field goal probabilities than do most professional sports commentators.

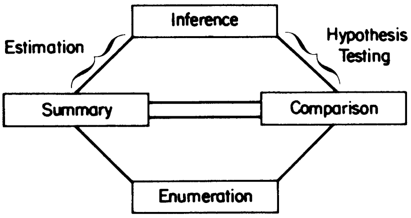


FIG. 2. *Four basic statistical operations, and how they relate to estimation.*

In order to look at a body of data it first must be summarized. *Summary* is the amalgamation of data, usually noisy data, to reveal interesting common features of the situation, for example, that the 50% success distance for field goals is about 40 yards. The circles in Figure 1 summarize Cockroft's kicking record as five probabilities. The solid curve is a more drastic summary, in terms of a two-parameter probability model.

Comparison is the process opposite to summary, the pulling apart of a data set to reveal interesting differences. For example, it seems harder to kick a field goal from 55 years than from 25 yards. Cockroft seems to be less successful than the AFC from 35–45 yards.

Much of our training as mathematical statisticians focuses on the top box in Figure 2. Statistical *inference* extrapolates from the data at hand to what might be reasonably expected given a much large, perhaps infinite, data set. For example, ".234 ± .038" in the third field goal model means the following: if we had available say 10,000 rather than 124 attempts from 55 yards, we expect that the observed proportion of successes would be within .038 of .234, with about 68% confidence, in the usual terminology. An hypothesis test is an inference about whether or not an observed comparison would stand up given a lot more data. In this paper we shall not discuss the hypothesis testing side of Figure 2.

A typical data analysis is an interplay between summaries and comparisons, with the data being repeatedly pulled apart and recombined to reveal similarities and differences. Inferential theory is used as a check on this process, to prevent, as far as possible, errors due to the limitations of the data set. Which of the four estimates would we prefer for γ_{55}? Standard hypothesis tests do not reject the logistic models, and also do not reject the hypothesis that Cockroft's kicking ability is no different than that of the entire AFL. This suggests estimate (4) of γ_{55}, .228 ± .020.

Estimation is the theory that concerns making summaries and inferences about summaries. The inferences are usually in the form of point and interval estimates. Textbook presentations of estimation tend to lose sight of the summary itself, in the excitement over how the summary is used to form, say, a uniform minimum variance unbiased estimate of some specific quantity. On the other hand, Fisher's original presentation of maximum likelihood stressed summarization. Fisher's claim, which we shall examine in the following sections, was that maximum likelihood is a superior method of data summarization, no matter what specific inferences may eventually be needed.

Maximum likelihood as used in practice is really two theories, one for summarization, the other for making specific point and interval estimates. (In the following sections, "estimation" and "estimate" will refer to this second process, that of making specific point and interval guesses for unknown parameters.) The two theories are described explicitly in the next section. Meanwhile, it is worth remembering that the real difficulty in most estimation problems lies in deciding which data is relevant to the quantity being estimated, not the specific form of the estimator. Can we really use all 1494 data points in estimating γ_{55}, as we do with estimate (4), or should we restrict attention to Cockroft's 100 kicks, in which case estimate (2) is preferred? Making such decisions, in this case trading off variance for bias in the estimate of γ_{55}, is the point of the summary-comparison-inference cycle described above. (This approach to model building is not above criticism. The last example in Section 8 illustrates some of its limitations and suggests, vaguely, the outlines of a more ambitious theory.) The summary aspect of maximum likelihood theory fits well into this cycle, which accounts for a good deal of maximum likelihood's popularity.

4. Maximum likelihood summarization and estimation. We now describe the two aspects of maximum likelihood, as a summary device, and as a method of providing specific point and interval estimates. Given a family of probability densities for **X**,

$$\mathscr{F} = \{f_\theta, \theta \in \Theta\}, \tag{4.1}$$

we observe data $\mathbf{X} = \mathbf{x}$. Let $\hat\theta$ be that value of θ, assumed to exist, which maximizes the probability density $f_\theta(\mathbf{x})$. The *maximum likelihood summary* of the data, abbreviated MLS, is the density function corresponding to $\theta = \hat\theta$,

$$\text{MLS}: \hat f = f_{\hat\theta}. \tag{4.2}$$

There are three crucial points here. (i) The parameter "θ" as used here is only a name, and plays no role in the summarization process. Any other way of naming the members of $\mathscr{F}$ results in the same MLS $\hat f$, given the same data **x**. (ii) The MLS is not a number or a vector, it is a probability density. We are summarizing a data set by a probability mechanism. This will be particularly important in Section 5. (iii) The MLS is a superior method of data summarization. This was Fisher's main point, and will be examined in Section 6.

Next, suppose that $\gamma(f)$ is a parameter (function of the unknown probabilty mechanism) we wish to estimate. The *maximum likelihood estimate*, MLE, is the corresponding function of $\hat f$,

$$\text{MLE}: \hat\gamma = \gamma(\hat f). \tag{4.3}$$

Maximum likelihood estimation takes the maximum likelihood summary as being the true probability mechanism, and simply reads off any parameter of interest from the MLS.

The criticisms of maximum likelihood by Berkson (1980) and discussants are mostly criticisms of the MLE as a point estimator, not of the MLS as a summarizer. We will try to make this distinction explicit in what follows.

As an example, consider observing n replicates from a normal distribution with unknown mean θ but known variance σ^2, $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} \mathscr{N}(\theta, \sigma^2)$. We can write the family $\mathscr{F}$ as

$$\mathscr{F} = \{f_\theta = \mathscr{N}(\theta, \sigma^2)^n, \theta \in \mathscr{R}^1\}, \tag{4.4}$$

where the power notation is shorthand for the distribution on $\mathscr{R}^n$ given by the product of n independent $\mathscr{N}(\theta, \sigma^2)$ distributions. Having observed $\mathbf{X} = \mathbf{x}$, the MLS is $\hat f = \mathscr{N}(\bar x, \sigma^2)^n$, where $\bar x = \Sigma x_i/n$. For the parameter $\gamma(f) = e^\theta$, the MLE is $\hat\gamma = e^{\bar x}$.

In this case one might well prefer the estimate $\bar\gamma = e^{\bar x - \sigma^2/2n}$, which is uniform minimum variance unbiased for γ. For $\sigma^2 = 1$, $n = 10$ the ratio of expected squared errors is $E(\hat\gamma - \gamma)^2/E(\bar\gamma - \gamma)^2 = 1.13$. On the other hand, the MLS is unassailable as a summarizer, since knowing $\hat f$ is equivalent to knowing the sufficient statistic $\bar x$. One of Fisher's main points was that in cases such as this where a sufficient statistic exists, it is automatically captured by the MLS. Sufficiency was one of many ingenious evasions Fisher used to avoid a detailed quantitative theory of point and interval estimation (i.e., decision theory), some others being asymptotic efficiency, second order efficiency, pivotal quantities, and ancillarity.

What is the difference between a summary and an estimate? A summary may be used for comparative purposes, such as Cockroft versus the rest of the AFC, with no specific estimation problem in mind. (Comparison of summaries is often used to suggest which estimates and hypothesis tests are of interest.) A summary may be used qualitatively, "The probability of successfully kicking a field goal decreases smoothly with distance," with no attempt at quantitative assessment. Different functions of the same summary may be used to estimate the same parameter, e.g., $e^{\bar{x}}$ or $e^{\bar{x}-\sigma^2/2n}$ for estimating e^θ in the example above. A full summary such as the MLS is universal, applying to all questions we might subsequently ask about the situation, while an estimate will apply to a particular aspect of interest.

The virtue of a summary is that it gives the statistician a chance to look at the data in compact form before proceeding further with the analysis. It is much easier to compare Cockroft with the AFL using the logistic curves than in terms of the original 1494 data points.

It would be easy to confuse summaries with *descriptive statistics.* A descriptive statistic is purely a device for describing data already seen by the statistician, as opposed to an estimate, which relates to data that might plausibly be seen in the future. A summary, as we are using that term, occupies middle ground. It both describes the data already at hand, and is a large first step toward making specific estimates. Descriptive statistics, as explained in Tukey (1979) for example, can operate effectively without probability models for the data. On the other hand, the MLS is a device for going from the data *and* a family of probability models to an efficient summary.

Fisher's claim for the superiority of the maximum likelihood summary as presented in Section 6 has never been seriously challenged, and gives a solid theoretical basis to the popularity of maximum likelihood in applied work. This aura of superiority has transformed itself to maximum likelihood estimation, where it is less well founded. The MLE can be a useful estimation device, with definite limitations which we shall try to make clear in later sections.

5. Fisher's information bound. If maximum likelihood estimation acts as if the maximum likelihood summary is exactly true, how can the theory provide an assessment of estimation error? The answer is simple, but ingenious. Define

$$\varepsilon(f) = Sd_f(\hat{\gamma}), \tag{5.1}$$

the standard deviation of the MLE $\hat{\gamma} = \gamma(\hat{f})$ under the true probability mechanism f. Then the parameter ε can itself be estimated by maximum likelihood, $\hat{\varepsilon} = \varepsilon(\hat{f})$. Notice that we are using the fact that the MLS $\hat{f}$ is a probability mechanism, and not just a number or vector.

The standard error estimate $\hat{\varepsilon}$ is hard to compute in most cases. Fisher provided a famous approximation, the *Fisher information bound,*

$$\hat{\varepsilon} \approx 1/\sqrt{\mathscr{I}_{\hat{\theta}}(\gamma)}, \tag{5.2}$$

where $\mathscr{I}_\theta(\gamma)$ is the Fisher information for γ, evaluated at point θ in Θ. This is not the usual presentation of the information bound, and we will indicate the derivation of (5.2) in Section 6. However, the important point is that *the Fisher information bound is, approximately, the standard deviation of the MLE $\hat{\gamma}$, assuming that the MLS $\hat{f}$ is true.* (More precisely, the Fisher information bound approximates the *unconditional* standard deviation of $\hat{\gamma}$, averaged over the whole sample space. See the last two paragraphs of Section 6.)

For example, consider estimate (4) of Section 2, i.e. .228 ± .020. The "± .020" can be interpreted as follows. Let $\mathbf{X}^*$ be a hypothetical data set comprised of 1494 data points (y_i, Z_i^*), $i = 1, \cdots, 1494$. For 128 of the data points $y_i = 55$, for 449 of the data points $y_i = 45$, etc. as indicated in Figure 1. Each Z_i^* is an independent Bernoulli trial generated according to (2.1) with $(\alpha, \beta) = (83, -.082)$. Let $\hat{\gamma}_{55}^*$ be the probability of success at $y = 55$ obtained by fitting model (2.1) to the hypothetical data $\mathbf{X}^*$. Then .020 is, approximately, the standard deviation of $\hat{\gamma}_{55}^*$.

The star notation indicates that we are not dealing with the real data $\mathbf{X}$, which is generated according to the real though unknown probability mechanism f, but rather with hypothetical data $\mathbf{X}^*$ from the MLS mechanism $\hat{f}$. Since everything is known to the statistician about $\hat{f}$, the hypothetical data $\mathbf{X}^*$ can actually be generated using Monte Carlo methods. Suppose we independently generate $\mathbf{X}^*(1), \mathbf{X}^*(2), \cdots, \mathbf{X}^*(B)$, each $\mathbf{X}^*(j)$ consisting of 1494 points as described above, and calculate the corresponding γ_{55} estimates $\hat{\gamma}^*_{55}(1), \hat{\gamma}^*_{55}(2), \cdots, \hat{\gamma}^*_{55}(B)$. If the number of replications B is fairly large, we can estimate $\hat{\varepsilon}$ from

$$\hat{\varepsilon} = \{\textstyle\sum_{j=1}^{B}[\hat{\gamma}^*_{55}(j) - \hat{\gamma}^*_{55}(\cdot)]^2/(B-1)\}^{1/2}, \tag{5.3}$$

$\hat{\gamma}^*_{55}(\cdot) = \sum_j \hat{\gamma}^*_{55}(j)/B$, without using Fisher's approxmation (5.2). As $B \to \infty$, (5.3) approaches the original definition $\hat{\varepsilon} = \varepsilon(\hat{f})$. The author (Efron, 1979) has suggested using approach (5.3) under the name "bootstrap" for nonparametric situations, where the information bound is difficult to calculate and/or unreliable.

6. The geometry of maximum likelihood summarization. In this section we consider a particularly simple situation which illustrates the salient aspects of maximum likelihood summarization. Fisher's claim of superiority for the MLS will be made explicit. This section is more technical than the remainder of the paper.

We consider a finite sample space $\mathscr{X} = \{1, 2, \cdots, L\}$, so that a probability distribution on $\mathscr{X}$ is a vector $\pi = (\pi_1, \pi_2, \cdots, \pi_L)'$,

$$\pi_\ell = \text{Prob}\{X = \ell\}, \quad \ell = 1, 2, \cdots, L. \tag{6.1}$$

The vector π must lie in $\mathscr{S}_L$, the L dimensional simplex,

$$\mathscr{S}_L = \{\pi : \pi_\ell \geq 0, \textstyle\sum_{\ell=1}^{L} \pi_\ell = 1\}. \tag{6.2}$$

We restrict the choice further by assuming that π belongs to a subset $\mathscr{F}_1$ of $\mathscr{S}_L$ indexed by a real parameter θ,

$$\mathscr{F}_1 = \{\pi(\theta), \theta \in \Theta\}, \tag{6.3}$$

Θ a possibility infinite interval of the real line.

The data vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ consists of n i.i.d. replicates from $\pi(\theta)$. The observed data $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ has density function

$$f_\theta(\mathbf{x}) = \textstyle\prod_{i=1}^{n} \pi_{x_i}(\theta). \tag{6.4}$$

Using the power notation of (4.4), the family of probability models for $\mathbf{X}$ is

$$\mathscr{F} = \{f_\theta = \pi(\theta)^n, \theta \in \Theta\}. \tag{6.5}$$

Let $\mathbf{p} = (p_1, p_2, \cdots, p_L)'$ be the proportions of the observations x_i in the L categories,

$$p_\ell = \#\{x_i = \ell\}/n. \tag{6.6}$$

Define $\eta_\ell(\theta) = \log \pi_\ell(\theta)$ and $\dot{\eta}_\ell(\theta) = \frac{\partial}{\partial\theta}\eta_\ell(\theta)$, so

$$\dot{\boldsymbol{\eta}}(\theta) = (\dot{\pi}_1(\theta)/\pi_1(\theta), \cdots, \dot{\pi}_L(\theta)/\pi_L(\theta)), \tag{6.7}$$

the dot indicating differentiation with respect to θ. Using (6.4), the function $\partial/\partial\theta \log f_\theta(\mathbf{x})$ equals

$$n \textstyle\sum_{\ell=1}^{L} \{p_\ell - \pi_\ell(\theta)\}\dot{\eta}_\ell(\theta) = n\{\mathbf{p} - \pi(\theta)\}'\dot{\boldsymbol{\eta}}(\theta). \tag{6.8}$$

From this, we see that the linear subset $\mathscr{L}(\hat{\theta})$ of $\mathscr{S}_L$ passing through $\pi(\hat{\theta})$ orthogonal to $\dot{\boldsymbol{\eta}}(\hat{\theta})$,

$$\mathscr{L}(\hat{\theta}) = [\mathbf{p} : \{\mathbf{p} - \pi(\hat{\theta})\}'\dot{\boldsymbol{\eta}}(\hat{\theta}) = 0], \tag{6.9}$$

corresponds to those data vectors $\mathbf{x}$ having $\hat{\theta}$ as a solution to the maximum likelihood equations $(\partial/\partial\theta) \log f_\theta(\mathbf{x}) = 0$. A careful discussion of the geometry of maximum likelihood

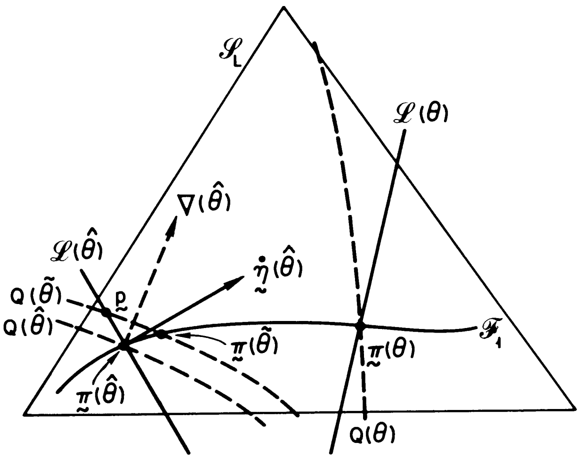

$\mathscr{S}_L$
$\mathscr{L}(\theta)$
$\nabla(\hat{\theta})$
$\mathscr{L}(\hat{\theta})$
$\dot{\boldsymbol{\eta}}(\hat{\theta})$
$Q(\tilde{\theta})$
$\mathbf{p}$
$Q(\hat{\theta})$
$\boldsymbol{\pi}(\tilde{\theta})$
$\boldsymbol{\pi}(\theta)$
$\mathscr{F}_1$
$\boldsymbol{\pi}(\hat{\theta})$
$Q(\theta)$

FIG. 3. *Maximum likelihood summarization*: $\mathscr{L}(\hat{\theta})$ *is the level surface of data vectors* $\mathbf{p}$ *having MLS* $\hat{f} = \pi(\hat{\theta})^n$. *Another summarization method, with level surface* $Q(\tilde{\theta})$, *is also indicated. The vector* $\dot{\boldsymbol{\eta}}(\theta)$ *has* ℓth *component* $\dot{\pi}_\ell(\theta)/\pi_\ell(\theta)$.

summarization, including the problems of multiple solutions, second-order efficiency, statistical curvature, and Kullback-Leibler distance, can be found in Efron (1975b, 1978).

We can now give a simple picture of maximum likelihood summarization: *data vectors* $\mathbf{x}$ *having* $\mathbf{p}$ *on the level surface* $\mathscr{L}(\hat{\theta})$ *are summarized by the same probability distribution,* $\hat{f} = \pi(\hat{\theta})^n$. This is illustrated in Figure 3.

What is so good about the MLS? Fisher's original answer was in terms of *asymptotic efficiency*. It simplifies discussion to think of $\hat{\theta}$ as an estimate of θ, even though we will eventually interpret the results in terms of summary rather than estimation. An observed vector of proportions $\mathbf{p}$ determines $\hat{\theta} = \hat{\theta}(\mathbf{p})$ by projection to $\mathscr{F}_1$ along the level surfaces $\mathscr{L}$ as indicated in Figure 3. Let $\tilde{\theta}$ be another way of estimating θ, comparable to $\hat{\theta}$ in two ways: $\tilde{\theta}$ is a function of $\mathbf{x}$ through $\mathbf{p}$, with the function $\tilde{\theta}(\mathbf{p})$ being defined for all $\mathbf{p} \in \mathscr{S}_L$, and

$$\tilde{\theta}\{\pi(\theta)\} = \theta. \tag{6.10}$$

Notice that $\hat{\theta}$ satisfies (6.10).

The first condition is easily justified since $\mathbf{p}$ is a sufficient statistic. Condition (6.10), called "*Fisher consistency*," says that $\tilde{\theta}$ and $\hat{\theta}$ are estimating the same thing, in the sense that if $\mathbf{p}$ falls on $\mathscr{F}_1$, then both $\tilde{\theta}$ and $\hat{\theta}$ give the same numerical estimate. This restriction on the form of $\tilde{\theta}$ also turns out to be innocuous for summary, though not for estimation, as discussed later in the section.

Let $Q(\theta)$ be the level curves of constant estimation using $\tilde{\theta}$,

$$Q(\theta) = \{\mathbf{p}: \tilde{\theta}(\mathbf{p}) = \theta\}. \tag{6.11}$$

The theory of *first order efficiency* concerns what happens if, as in Figure 3, $Q(\theta)$ is not tangent to $\mathscr{L}(\theta)$ at the point of intersection with $\mathscr{F}_1$. Let $\nabla(\theta)$ be the orthogonal to $Q(\theta)$ at the point $\pi(\theta)$, and define

$$\cos^2 A_\theta = \frac{\{\dot{\boldsymbol{\eta}}(\theta)'\boldsymbol{\Sigma}_\theta \nabla(\theta)\}^2}{\{\dot{\boldsymbol{\eta}}(\theta)'\boldsymbol{\Sigma}_\theta \dot{\boldsymbol{\eta}}(\theta)\}\{\nabla(\theta)'\boldsymbol{\Sigma}_\theta \nabla(\theta)\}}, \tag{6.12}$$

the squared cosine of the angle between $\dot{\boldsymbol{\eta}}(\theta)$ and $\nabla(\theta)$, in the inner product determined by the matrix $\boldsymbol{\Sigma}_\theta$ with ℓmth element $\pi_\ell(1 - \pi_\ell)$ if $\ell = m$, $-\pi_\ell \pi_m$ if $\ell \neq m$.

Under mild regularity conditions (Rao, 1973, Section 5e),

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathscr{N}\left(0, \frac{1}{i_\theta}\right), \tag{6.13}$$

where $i_\theta = \sum_{\ell=1}^{L} \dot{\pi}_\ell(\theta)^2/\pi_\ell(\theta)$ is the Fisher information in a single observation, and also

$$\sqrt{n}(\tilde{\theta} - \theta) \rightarrow \mathcal{N}\left(0, \frac{1}{i_\theta \cos^2 A_\theta}\right). \tag{6.14}$$

This shows that $\tilde{\theta}$ is asymptotically inferior to $\hat{\theta}$ unless $A_\theta = 0$, that is unless $Q(\theta)$ is tangent to $\mathcal{L}(\theta)$ at $\pi(\theta)$. Fisher liked to state this in terms of relative sample sizes: $\tilde{\theta}$ based on n observations has the same asymptotic distribution as $\hat{\theta}$ based on $n \cos^2 A_\theta$ observations. Using $\tilde{\theta}$ instead of $\hat{\theta}$ wastes proportion $\sin^2 A_\theta$ of the observations. This makes it clear that the superiority of $\hat{\theta}$ is not tied to any specific estimation problem.

In fact, $\tilde{\theta}$ can be thought of, asymptotically, as $\hat{\theta}$ plus random noise. Working in "statistically large" neighborhoods of some fixed value θ_0, say $\theta \in \theta_0 \pm n^{-1/3}$, we have

$$\tilde{\theta} = \hat{\theta} + \frac{\tan A_{\theta_0}}{\sqrt{n i_{\theta_0}}} Z + o_p(n^{-1/2}), \tag{6.15}$$

where $Z \sim \mathcal{N}(0, 1)$ is independent of $\hat{\theta}$ (Efron, 1975b). A statistician given $\hat{\theta}(\mathbf{p})$ and a random number table can, asymptotically at least, duplicate the performance of any decision rule based on $\tilde{\theta}(\mathbf{p})$. From the decision-theoretic viewpoint, asymptotic sufficiency is a more accurate name than asymptotic efficiency. Result (6.15) is an early example of an asymptotic complete-class theorem.

If $\mathcal{F}$ is a one-parameter exponential family, then the summary surfaces $\mathcal{L}(\theta)$ are parallel to each other, and $\hat{\theta}$ is a genuine sufficient statistic. If not, we can approximate $\mathcal{F}$ near any parameter value θ_0 by the one-parameter exponential family $\mathcal{F}_0$ having log probability vector $\eta(\theta_0) + \dot{\eta}(\theta_0)(\theta - \theta_0)$. Then the level surfaces for $\mathcal{F}$ and $\mathcal{F}_0$ will agree at $\theta = \theta_0$, both being $\mathcal{L}(\theta_0)$ there. In this sense the MLS is *locally sufficient* as well as asymptotically sufficient.

Sufficiency says everything about summarization, but leaves open the choice of specific estimates for specific estimation problems. In the example of Section 4, we certainly want to estimate $\gamma = e^\theta$ with a function of $\bar{x}$, but not necessarily the obvious function $e^{\bar{x}}$. The MLE is a simple way to use the MLS for estimation. It does *not* enjoy the same degree of theoretical justification as does the MLS. See Remark D, Section 7.

Pearson's *method of moments*, which preceded Fisher's theory, has $Q(\theta)$ linear, but $A_\theta \neq 0$, in most cases. Other methods, such as *minimum chi squared*, satisfy the tangency property $A_\theta \equiv 0$, but have $Q(\theta)$ curved instead of linear. *Second order efficiency* describes a more delicate version of (6.15) appropriate to this case,

$$\tilde{\theta} = \hat{\theta} + \frac{1}{n} \sum_{j=1}^{L-2} \lambda_j(\theta_0) Z_j^2 + o_p(n^{-1}), \tag{6.16}$$

where the $\lambda_j(\theta)$ are known functions of θ, and $Z_j \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, independent of $\hat{\theta}$. See Efron (1975b), Rao (1962), Ghosh and Subramanyan (1974), and Pfanzagl and Wefelmeyer (1978) for discussions of second order efficiency.

The fact that the $\mathcal{L}(\theta)$ are linear has an appealing consequence in terms of combining independent experiments. Suppose we take independent samples of size n_1 and n_2 according to density (6.4), θ the same in both samples, and observe vectors of proportions $\mathbf{p}_1$ and $\mathbf{p}_2$ respectively. Suppose also that both $\mathbf{p}_1$ and $\mathbf{p}_2$ lie on the same level surface $\mathcal{L}(\hat{\theta})$. Then so will the linear combination $\mathbf{p} = (n_1\mathbf{p}_1 + n_2\mathbf{p}_2)/(n_1 + n_2)$. In other words, if both data sets have MLS corresponding to $\theta = \hat{\theta}$, then so will the combined data set. This seemingly obvious rule of combination is violated by any summary method having curved level surfaces.

Fisher consistency, (6.10), seems to stack the deck in favor of maximum likelihood since we insist that $\tilde{\theta}$ have a property which is already known to hold for $\hat{\theta}$. In this discussion, however, "θ" is not a genuine parameter of interest, only a convenient way to name the distributions in $\mathcal{F}$. If $\tilde{\theta}(\mathbf{p})$ is any statistic defined for all $\mathbf{p} \in \mathcal{S}_L$, we can *define* the name θ by $\theta \equiv \tilde{\theta}(\pi)$ for $\pi \in \mathcal{F}_1$. The asymptotic sufficiency of $\hat{\theta}$, (6.15) or (6.16), still holds.

If θ is a genuine parameter of interest which we want to estimate, then renaming $\mathcal{F}$, as above, is not so innocuous. This again reflects the difference between summary and more specific estimation problems. Fisher consistency is not necessarily desirable in the latter context. Notice that in the example of Section 4, the estimate $\bar{\gamma} = e^{\bar{x}-\sigma^2/2n}$ is the MLE of $e^{\theta-\sigma^2/2n}$, so $\bar{\gamma}$ is Fisher consistent for this parameter rather than for the parameter of interest e^{θ}.

It is easy to motivate approximation (5.2), the Fisher information bound, in terms of Figure 3. We take the parameter $\gamma(f)$ to be θ itself. Let $\mathbf{X}^*$ be the hypothetical data vector drawn according to the MLS $\hat{f}$, and $\mathbf{p}^*$ the corresponding vector of proportions, $p_\ell^* = \#\{X_i^* = \ell\}/n$. In this situation $\mathbf{p}^*$ is distributed according to the rescaled multinomial distribution on L categories, n draws, probability vector $\pi(\hat{\theta})$,

$$\mathbf{p}^* \sim \text{Mult}(n, \pi(\hat{\theta}))/n. \tag{6.17}$$

We show below that $\hat{\theta}^* = \hat{\theta}(\mathbf{p}^*)$ has the first order Taylor expansion

$$\hat{\theta}(\mathbf{p}^*) = \hat{\theta} + \frac{\dot{\eta}(\hat{\theta})'\{\mathbf{p}^* - \pi(\hat{\theta})\}}{i_{\hat{\theta}}} \tag{6.18}$$

as a function of $\mathbf{p}^*$, with $\hat{\theta}$ fixed. This gives the approximation

$$\text{Var}\,\hat{\theta}(\mathbf{p}^*) = \frac{\dot{\eta}(\hat{\theta})'\text{Cov}(\mathbf{p}^*)\dot{\eta}(\hat{\theta})}{i_{\hat{\theta}}^2}, \tag{6.19}$$

where $\text{Cov}(\mathbf{p}^*)$ is the covariance matrix of $\mathbf{p}^*$ under (6.17), having ℓmth entry $\pi_\ell(\hat{\theta})(1 - \pi_\ell(\hat{\theta}))/n$ for $\ell = m$, and $-\pi_\ell(\hat{\theta})\pi_m(\hat{\theta})/n$ for $\ell \neq m$. However $\dot{\eta}(\hat{\theta})'\,\text{Cov}(\mathbf{p}^*)\dot{\eta}(\hat{\theta}) = i_{\hat{\theta}}/n$ (using $\Sigma\dot{\pi}_\ell(\hat{\theta}) = (\partial/\partial\theta)\Sigma\pi_\ell(\theta)|_{\hat{\theta}} = 0$), so (6.19) gives (5.2),

$$\hat{\varepsilon}^2 = \text{Var}\,\hat{\theta}(\mathbf{p}^*) \doteq \frac{1}{n i_{\hat{\theta}}} = \frac{1}{\mathscr{I}_{\hat{\theta}}}, \tag{6.20}$$

where $\mathscr{I}_{\hat{\theta}} = n i_{\hat{\theta}}$ is the total information for estimating θ, evaluated at $\theta = \hat{\theta}$.

It remains to verify (6.18). Let $d\hat{\theta}$ indicate the change in the MLE $\hat{\theta}$ resulting from change $\mathbf{dp}$ in $\mathbf{p}$. Since $\hat{\theta} + d\hat{\theta}$ must satisfy the likelihood equation for data vector $\mathbf{p} + \mathbf{dp}$, (6.9) gives

$$0 = \dot{\eta}(\hat{\theta})'\mathbf{dp} - [\dot{\eta}(\hat{\theta})'\dot{\pi}(\hat{\theta}) - \{\mathbf{p} - \pi(\hat{\theta})\}'\ddot{\eta}(\hat{\theta})]\,d\hat{\theta}, \tag{6.21}$$

ignoring higher order differentials. Notice that $\dot{\eta}(\hat{\theta})'\dot{\pi}(\hat{\theta}) = i_{\hat{\theta}}$. *Setting* $\mathbf{p} = \pi(\hat{\theta})$ and $\mathbf{p}^* = \pi(\hat{\theta}) + \mathbf{dp}$ gives (6.18). This result appears in Jaeckel (1971).

Higher order summaries. Fisher (1934) argued persuasively that in some circumstances the *observed Fisher information*

$$I(\mathbf{x}) = \frac{\partial^2}{\partial\theta^2}\log f_\theta(\mathbf{x})\Big|_{\theta=\hat{\theta}}$$

is superior to $\mathscr{I}_{\hat{\theta}}$ as a measure of information, in the sense that $1/I(\mathbf{x})$ is a more relevant estimate of the variance of $\hat{\theta}$. See Efron and Hinkley (1978) for a discussion and extension of Fisher's results, and some dramatic numerical examples of their practical importance. Usually $I(\mathbf{x})$ is not recoverable from the MLS $\hat{f}$, or equivalently $\hat{\theta}$, so that we have to retain the bigger summary $(\hat{\theta}, I(\mathbf{x}))$ in order to make more exact inferences.

There is no contradiction here with the previous assertion that $\hat{\theta}$ is a superior summary device. If we want the most compact possible summary of the data, just one number in the context of Figure 3, then $\hat{\theta}$ is indeed asymptotically superior. By summarizing the data less succinctly, keeping track of $I(\mathbf{x})$ as well as $\hat{\theta}$, we can make better inferences. This line of thought ends with the entire likelihood function as the summary statistic. The likelihood function is always sufficient, but is too clumsy to be an effective summary device in most situations, especially when θ is multidimensional. The summarization process usually stops with $\hat{\theta}$, equivalently $\hat{f}$, and either the information matrix $\mathscr{I}_{\hat{\theta}}$ or the matrix of second derivatives of the log likelihood function evaluated at $\hat{\theta}$.

Invariance. The MLS behaves correctly under invariant transformations of $\mathscr{F}$. Suppose there exists one-to-one transformations mapping the parameter and sample spaces onto themselves, $\theta' = g(\theta)$ and $\mathbf{x}' = h(\mathbf{x})$, which leave $\mathscr{F}$ invariant. That is,

$$f^{\mathbf{X}'}_{\theta'}(\mathbf{x}') = f_\theta(\mathbf{x})\, J(\mathbf{x} \to \mathbf{x}') = f_{\theta'}(\mathbf{x}'), \tag{6.22}$$

where $J(x \to x')$ is the Jacobian. The last equality gives $\hat{\theta}(\mathbf{x}') = g(\hat{\theta}\,(\mathbf{x}))$, or equivalently $g\hat{\theta}(h^{-1}\mathbf{x}') = \hat{\theta}\,(\mathbf{x}')$, which shows that the MLS maps invariantly.

Minimum Distance Methods. Let $D(x, \pi) \equiv -\log \pi_x$ be thought of as a measure of discrepancy between the possible outcome $X = x$ and the probability vector π. Since

$$\textstyle\sum_{i=1}^{n} D(x_i, \pi(\theta)) = -n \sum_{\ell=1}^{L} p_\ell \log \pi_\ell(\theta) = -\log \sum_{\ell=1}^{L} \pi_\ell(\theta)^{np_\ell}, \tag{6.23}$$

we see that minimizing $\Sigma_i D(x_i, \pi(\theta))$ as a function of θ is the same as maximizing the likelihood. Maximum likelihood summary can be thought of as a *minimum distance method* (Wolfowitz, 1957), defined in terms of the component observations x_i.

The fact that maximum likelihood can be defined in terms of the individual summands $D(x_i, \pi(\theta))$ is a handy property. In situations like logistic regression with continuous prediction variables, which the football example would be if we could observe the original ungrouped data, it means that preliminary grouping is not necessary to compute the MLS; see Efron (1978b). Any choice of $D(x, \pi)$ other than $-\log \pi_x$ makes the minimizer of $\Sigma_i D(x_i, \pi(\theta))$ first order inefficient; the level curves will be linear, but different from the $\mathscr{L}(\theta)$.

Robust estimation. The theory of robust estimation, as described in Huber (1981), considers the possibility of error in the specification of $\mathscr{F}$. The MLS theory, as pictured in Figure 3, is modified in the following way: the level surfaces of equivalent summary are kept linear, at least in m-estimation, but the orthogonal vector is constrained never to point too closely toward a corner of $\mathscr{S}_L$. In other words, the summary is not allowed to be overly influenced by any one observation. This agrees with the characterization of statistics as the science of gathering information in small pieces, as discussed in Section 3. (In his discussion of Berkson's paper (1980), LeCam, expanding on a result of Bahadur's, gives an interesting example where some robustification is necessary to make the MLS work, even assuming that $\mathscr{F}$ is described correctly.

Robustness is an important addition to maximum likelihood theory. Whether it applies to summary or estimation, or both, is still a point of debate. See Pratt's comments on Stigler's paper (1977) for an incisive discussion. The robustness literature is mainly decision-theoretic in character, and represents a potentially major Waldian contribution to the practice of statistical estimation. Robustifying maximum likelihood in a way which retains the convenience and generally high efficiency of Fisher's theory is a formidable project. Good progress has been made in robust regression and a few other areas; see Huber (1981).

7. Some remarks on the MLE.

Remark A. Suppose θ increases monotonically as we move from left to right along $\mathscr{F}_1$ in Figure 3. The MLE $\hat{\theta}$ will be smaller or greater than the true value θ as $\mathbf{p}$ falls to the left or right of $\mathscr{L}(\theta)$, since $\hat{\theta}\,(\pi(\theta)) = \theta$, because of Fisher consistency. But $\mathbf{p} \sim \text{Mult}(n, \pi(\theta))/n$ has a limiting normal distribution centered at $\pi(\theta)$, so that

$$\text{Prob}\{\hat{\theta} < \theta\} = .50 + O(n^{-1/2}). \tag{7.1}$$

In general, asymptotic *median unbiasedness* to order $n^{-1/2}$ holds for the MLE under reasonable regularity conditions. The same result to $O(n^{-1})$ is true for unbiasedness in the usual expectation sense; see Cox and Hinkley (1974, page 310). Unbiasedness is a popular property among consumers of statistical methods since it conveys a feeling of scientific objectivity. This point comes up again in Section 8, where we discuss estimates which are deliberately biased.

Remark B. Once we have calculated the MLS $\hat{f}$, we have available the MLE for every possible parameter $\gamma(f)$. The automatic way in which it produces estimates for even very complicated parameters is another popular feature of maximum likelihood estimation. This is particularly true since modern computers have made MLE computations feasible in most situations.

The fact that it automatically estimates all possible parameters strongly suggests that the MLE can be non-optimal if the statistician has one specific estimation problem in mind. Arbitrarily bad counterexamples, along the line of the e^{θ} example in Section 4, are easy to construct. Nevertheless the MLE has a good reputation, acquired over 60 years of heavy use, for producing reasonable point estimates. Useful general improvements on the MLE, such as robust estimation (Section 6) and Stein estimation (Section 8), are all the more impressive for their rarity.

Remark C. Uniform minimum variance unbiased estimation is perhaps the best known competitor to the MLE. Its major defect relates to Remark B: unlike the MLE, it is difficult or impossible to produce UMVU estimates in most situations. (Imagine how popular UMVU estimates would be if they could always be produced as easily as the MLE.)

Bayesian estimation, most often using a squared error loss function, is another competitor just as universal as the MLE approach. A perceived lack of objectivity, Remark A, seems to be the main barrier to the routine use of Bayesian methods. Efforts to construct an objective Bayesian theory along the lines of Jeffreys (1967) have proved unexpectedly difficult; see Efron (1978c).

Remark D. How do the asymptotics of Section 6 relate to point estimation? If $\tilde{\theta}$ is Fisher consistent for θ, but first-order inefficient, then (6.13) and (6.14) show that $\tilde{\theta}$ is asymptotically inferior to $\hat{\theta}$ as a point estimator of θ, for any reasonable loss function.

If $\tilde{\theta}$ is first-order efficient, then a more careful statement is necessary: some function of $\hat{\theta}$, not necessarily $\hat{\theta}$ itself, will asymptotically dominate $\tilde{\theta}$. Ghosh, Sinha, and Wieand (1980) gave a nice discussion and proof. Their approach can roughly be described as follows: let $\beta(\theta) = E_{\theta}(\tilde{\theta} - \hat{\theta})$. Then $\bar{\theta} = \hat{\theta} + \beta(\hat{\theta})$ has the same expectation as $\tilde{\theta}$, and less variability, to a high order of approximation. It is $\bar{\theta}$ which asymptotically dominates $\tilde{\theta}$. Here we are using $\hat{\theta}$ to correct itself by estimating β, much as it was used to estimate ε in Section 5. Notice however, that this theory does not automatically produce a good estimate for θ. All that is said is how to find one as good or better than a given estimate $\tilde{\theta}$.

Remark E. The MLE "maps correctly" in the sense that if $\delta = g(\gamma)$, then $\hat{\delta} = g(\hat{\gamma})$. This mapping property will hold for any estimation method which first estimates the entire probability mechanism f, and then reads off parameter estimates as if the estimate of f were true. Minimum distance methods, mentioned in Section 6, are usually used in this way.

Remark F. Suppose $\mathbf{A}(\mathbf{X})$ is an *ancillary statistic*, one whose marginal distribution does not depend upon the unknown parameter θ, so that we can write the density function as

$$f_{\theta}(\mathbf{x}) = f^{\mathbf{A}}(\mathbf{a})\, f_{\theta}^{\mathbf{X}|\mathbf{A}}(\mathbf{x}\,|\,\mathbf{a}).$$

Notice that the MLE $\hat{\theta}$, the maximizer of $f_{\theta}(\mathbf{x})$, is also the maximizer of $f_{\theta}^{\mathbf{X}|\mathbf{A}}(\mathbf{x}\,|\,\mathbf{a})$. If

$\gamma(\theta)$ is any function of θ, then the MLE $\hat{\gamma} = \gamma(\hat{\theta})$ has the same value whether or not we condition on $\mathbf{A} = \mathbf{a}$. Here θ is assumed to have an intrinsic meaning which can be defined without reference to $\mathscr{F}$, and $\gamma(\theta)$ is a function of θ which does not depend on how θ relates to $\mathscr{F}$. In (4.4), for example, $\gamma = e^{\theta}$ is allowable but $\varepsilon = Sd(\hat{\gamma})$ is not, as the next paragraph shows.

Fisher (1934) argued persuasively that it is more relevant for statistical inference to consider the conditional family of distributions $\mathscr{F}(\mathbf{a}) = \{f_\theta^{\mathbf{X}|\mathbf{A}}(\mathbf{x}\,|\,\mathbf{a}),\ \theta \in \Theta\}$ than the unconditional family $\mathscr{F} = \{f_\theta(\mathbf{x}),\ \theta \in \Theta\}$. The MLE $\hat{\theta}$, or $\hat{\gamma}$ for any function $\gamma(\theta)$, is the same in either case, but the MLS is different: $f_{\hat{\theta}}^{\mathbf{X}|\mathbf{A}}(\mathbf{x}\,|\,\mathbf{a})$ versus $f_{\hat{\theta}}(\mathbf{x})$. (Hinkley (1981) makes the important point that the discarded factor $f^{\mathbf{A}}(\mathbf{a})$ is crucial for testing the adequacy of the family $\mathscr{F}$.)

This distinction concerning the MLS has important practical consequences. The standard deviation parameter ε is estimated differently at (5.1), depending on which MLS is used. Consider the case where θ is real-valued, and where the parameter of interest is θ itself, so that $\varepsilon = Sd(\hat{\theta})$. Efron and Hinkley (1978) and Hinkley (1980), show that approximation (5.2), $\hat{\varepsilon} \approx 1/\sqrt{\mathscr{I}_{\hat{\theta}}}$, appropriate for the unconditional family $\mathscr{F}$, is better replaced by

$$\hat{\varepsilon} \approx 1/\sqrt{I(\mathbf{x})}, \qquad I(\mathbf{x}) = -\frac{\partial^2}{\partial\theta^2}\log f_\theta(\mathbf{x})\,\bigg|_{\theta=\hat{\theta}}$$

in $\mathscr{F}(\mathbf{a})$.

Remark G. Suppose the observed data $\mathbf{x}$ were obtained by some form of sequential sampling. The football data, for example, might have been collected from the beginning of 1969 until the first time Cockroft kicked successfully from beyond 50 yards. As in Remark F, this does not affect the MLE $\hat{\theta}$, but does change the MLS.

Remark H. Suppose the data consists of two parts, say $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, but we lose $\mathbf{z}$. Let $\hat{\theta}(\mathbf{y})$ be the MLE based just on $\mathbf{y}$. We can imagine augmenting $\mathbf{y}$ with an infinite amount of artificial data generated according to $f_{\hat{\theta}(\mathbf{y})}(\mathbf{z}\,|\,\mathbf{y})$. If $\mathbf{z}$ were discrete, taking on K possible values, the artificial data would consist of values $(\mathbf{y}, \mathbf{z}_1), (\mathbf{y}, \mathbf{z}_2), \cdots, (\mathbf{y}, \mathbf{z}_K)$, with proportion $f_{\hat{\theta}(\mathbf{y})}(\mathbf{z}_k\,|\,\mathbf{y})$ of $(\mathbf{y}, \mathbf{z}_k)$. The maximum likelihood estimate of θ based on the artificial data set still equals $\hat{\theta}(\mathbf{y})$. (Proof below.) This *self-consistency* property of the MLE has been rediscovered in many different contexts since Fisher's original papers. Dempster, Rubin, and Laird (1976) make it the basis of their *EM algorithm* for calculating the MLE in missing data situations.

Proof of self-consistency. Let $f_\theta^{\mathbf{Y}}(\mathbf{y})$ be the marginal density of $\mathbf{y}$, and $\dot{f}_\theta^{\mathbf{Y}}(\mathbf{y}) = (\partial/\partial\theta)\, f_\theta^{\mathbf{Y}}(\mathbf{y})$. Then

$$0 = \frac{\dot{f}_{\hat{\theta}(\mathbf{y})}^{\mathbf{Y}}(\mathbf{y})}{f_{\hat{\theta}(\mathbf{y})}^{\mathbf{Y}}(\mathbf{y})} = \int \frac{\dot{f}_{\hat{\theta}}(\mathbf{y}, \mathbf{z})}{f_{\hat{\theta}(\mathbf{y})}(\mathbf{y}, \mathbf{z})}\, \frac{f_{\hat{\theta}(\mathbf{y})}(\mathbf{y}, \mathbf{z})}{f_{\hat{\theta}(\mathbf{y})}^{\mathbf{Y}}(\mathbf{y})}\, d\mathbf{z} = \int \frac{\dot{f}_{\hat{\theta}(\mathbf{y})}(\mathbf{y}, \mathbf{z})}{f_{\hat{\theta}(\mathbf{y})}(\mathbf{y}, \mathbf{z})}\, f_{\hat{\theta}(\mathbf{y})}(\mathbf{z}\,|\,\mathbf{y})\, d\mathbf{z}. \tag{7.2}$$

The last quantity in (7.2) equaling zero shows that $\hat{\theta}(\mathbf{y})$ satisfies the MLE equations for the artificial data set, pretending that the artificial data is an i.i.d. sample (of infinite size) from $f_\theta(\mathbf{y}, \mathbf{z})$.

Remarks E, F, G, and H show four invariance properties of the MLE. These are advantageous to the statistical practitioner, who gets the same estimate under a variety of changing circumstances and hence can worry less about the circumstances. On the other hand, worrying about the circumstances may give better estimates, as in the e^{θ} example or the baseball example of the next section.

8. The MLS, MLE, and Stein's phenomenon. The second column of Table 1, taken from Efron (1975a) and Efron and Morris (1975) shows the observed batting averages of 18 major league baseball players after their first 45 times at bat in 1970. We assume that each player's results are independent Bernoulli trials, with true probability θ_i of a hit for player i,

$$(8.1)\qquad X_{ij} = \begin{matrix} 1\ (\text{“}hit\text{”}) \\ \\ 0\ (\text{“}out\text{”}) \end{matrix} \quad \text{prob.} \quad \begin{matrix} \theta_i \\ \\ 1-\theta_i \end{matrix} \qquad \text{indep. } i = 1, 2, \cdots, 18, \quad j = 1, 2, \cdots, 45,$$

so that the MLE $\hat{\theta}_i = \Sigma_j X_{ij}/45$ has expectation θ_i and variance $\sigma_i^2 = \theta_i(1-\theta_i)/45$.

The maximum likelihood summary is (8.1), with $\hat{\theta}_i$ replacing θ_i,

$$(8.2)\qquad \text{MLS: } X_{ij}^* = \begin{matrix} 1 \\ \\ 0 \end{matrix} \quad \text{prob.} \quad \begin{matrix} \hat{\theta}_i \\ \\ 1-\hat{\theta}_i \end{matrix} \qquad \text{indep. } i = 1, \cdots, 18, \quad j = 1, \cdots, 45.$$

Knowing the MLS is equivalent to knowing the sufficient statistic $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_{18})$, so that the MLS perfectly summarizes the data in this case.

Stein (1955) pointed out a disturbing phenomenon of high dimensional maximum likelihood estimation. For the data in column 2 of Table 1, we calculate $\Sigma_{i=1}^{18}(\hat{\theta}_i - \hat{\theta}.)^2 = .083$. However, for hypothetical data calculated according to the MLS (8.2), defining $\hat{\theta}_i^* = \Sigma_j X_{ij}^*/45$ and $\hat{\theta}_\cdot^* = \Sigma_i \hat{\theta}_i^*/18$, it is easy to see that $E\Sigma_i(\hat{\theta}_i^* - \hat{\theta}_\cdot^*)^2 = .157$, nearly twice the observed value. The probability that $\Sigma_i(\hat{\theta}_i^* - \hat{\theta}_\cdot^*)^2$ will be equal or less than .083 is about .034. We can see that the MLS (8.2) is unlikely to have generated the type of data actually observed!

Stein's phenomenon suggests that it is dangerous to read a certain parameter γ from the MLS, namely $\gamma = \Sigma(\theta_i - \theta.)^2$. The MLE $\hat{\gamma} = \Sigma(\hat{\theta}_i - \hat{\theta}.)^2$ is strongly biased upwards[1], by amount $(1 - 1/k)\Sigma_{i=1}^{k}\sigma_i^2$, $k = 18$. James and Stein (1960) used this fact to construct an estimate $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \cdots, \tilde{\theta}_k)$ which, under normality assumptions, is always closer to $\boldsymbol{\theta}$ than is $\hat{\boldsymbol{\theta}}$ in terms of expected squared distance,

$$(8.3)\qquad \tilde{\theta}_i = \hat{\theta}. + \left\{1 - \frac{(k-3)\sigma^2}{\hat{\gamma}}\right\}(\hat{\theta}_i - \hat{\theta}.).$$

Column 4 of Table 1 shows the $\tilde{\theta}_i$, computed from (8.3) with $\sigma^2 = \hat{\theta}.(1 - \hat{\theta}.)/45 = .0043$. In the baseball example, we have good independent estimates of the true θ_i, and we can see the superiority of $\tilde{\boldsymbol{\theta}}$ over $\hat{\boldsymbol{\theta}}$: $\Sigma(\tilde{\theta}_i - \theta_i)^2/\Sigma(\hat{\theta}_i - \theta_i)^2 = .29$. Details of this calculation appear in Efron and Morris (1975), and Efron (1975a).

Usually the statistician is not interested in simultaneously estimating all the θ_i. A typical situation is that certain low dimensional functions of $\boldsymbol{\theta}$ are of particular interest, often linear contrasts $\gamma_c = \Sigma c_i\theta_i$, where $\Sigma c_i = 0$. (For instance in an ANOVA model, γ_c might be the difference between two main effects.) A disturbing question raised by Stein's phenomenon is which linear contrasts γ_c can safely be read off of the MLS, i.e., estimated by the MLE $\hat{\gamma}_c$? Table 2 illustrates two aspects of the answer.

In the left half of Table 2, $\gamma_c = \frac{1}{9}\Sigma_A\theta_i - \frac{1}{9}\Sigma_B\theta_i$, $A = \{1, 2, 3, 5, 6, 9, 13, 16, 17\}$ and $B = \{4, 7, 8, 10, 11, 12, 14, 15, 18\}$. We might think of A as a Treatment group and B as a Control group. In fact A consists of the nine best hitters, as measured by the true θ_i in column 3, and B the nine worst hitters. This choice is guaranteed to make the "treatment" effective, but since the data going into column 3 is independent of that in column 2, Table 1, we can still legitimately compare estimating γ_c by $\hat{\gamma}_c = \frac{1}{9}\Sigma_A\hat{\theta}_i - \frac{1}{9}\Sigma_B\hat{\theta}_i$ or by $\tilde{\gamma}_c = \frac{1}{9}\Sigma_A\tilde{\theta}_i - \frac{1}{9}\Sigma_B\tilde{\theta}_i$.

Table 2 shows the MLE estimating γ_c quite well in this case, $\hat{\gamma}_c = .047 \pm .031$ compared to the true value $\gamma_c = .056$. The James-Stein estimate $\tilde{\gamma}_c = .010 \pm .016$ is much too small.

[1] In the context of Remark A, Section 7, let $\gamma(\theta) = \theta^2$ and suppose that the true value of θ is near zero, in the sense that both events $\{\hat{\theta} > |\theta|\}$ and $\{\hat{\theta} < -|\theta|\}$ have substantial probability of occurrence. Then the argument of Remark A fails for $\hat{\gamma}$, and Prob $\{\hat{\gamma} > \gamma\}$ is substantially greater than 0.50. Essentially the same argument applies to $\hat{\gamma} = \Sigma(\hat{\theta}_i - \hat{\theta}.)^2$.

TABLE 1

1970 batting averages for 18 major league players after their first 45 at bats. Taken from Efron (1975a) and Efron and Morris (1975)

i	MLE $\hat{\theta}_i$	Parameter value θ_i	$\tilde{\theta}_i$ James-Stein estimator	At bats, remainder of 1970
1	0.400	0.346	0.293	367
2	0.378	0.298	0.289	426
3	0.356	0.276	0.284	521
4	0.333	0.221	0.279	276
5	0.311	0.273	0.275	418
6	0.311	0.270	0.275	467
7	0.289	0.263	0.270	586
8	0.267	0.210	0.265	138
9	0.244	0.269	0.261	510
10	0.244	0.230	0.261	200
11	0.222	0.264	0.256	277
12	0.222	0.256	0.256	270
13	0.222	0.304	0.256	434
14	0.222	0.264	0.256	538
15	0.222	0.226	0.256	186
16	0.200	0.285	0.251	558
17	0.178	0.319	0.247	405
18	0.156	0.200	0.242	70
Player Number	Batting average after 45 at bats $\hat{\theta}_{\cdot} = .265$	Batting average remainder of season	$\dfrac{\Sigma(\tilde{\theta}_i - \theta_i)^2}{\Sigma(\hat{\theta}_i - \theta_i)^2} = .29$	Average remainder at bats = 369.3

The right half of Table 2 cuts the other way. Here $\gamma_c = \tfrac{1}{9}\,\Sigma_A\,\theta_i - \tfrac{1}{9}\,\Sigma_B\,\theta_i$ where $A = \{1, 2, \cdots, 9\}$ and $B = \{10, 11, \cdots, 18\}$; that is, γ_c is the true difference in average between the nine players with the highest observed average in their first 45 at bats, and the nine players with the lowest observed averages. In this case the James-Stein estimate .023 ± .016 is near the true value .010, while the MLE .111 ± .031 greatly overestimates the true difference.

Table 2 illustrates a general effect: prechosen contrasts (and other low-dimensional functions of $\boldsymbol{\theta}$) are well-estimated by the MLE. Contrasts obtained by "data-snooping" are better estimated by the James-Stein method. For any contrast γ_c, we have $\tilde{\gamma}_c =$ (factor) $\hat{\gamma}_c$, where (factor) $= \{1 - (k-3)\sigma^2/\hat{\gamma}\}$ is always less than 1. Shrinking the apparent contrasts $\hat{\gamma}_c$ toward zero is based on a Bayesian argument, see Efron and Morris (1972). The Bayesian argument breaks down for prechosen contrasts, ones for which there is a priori belief that an interesting effect may be present. This fact is noted in the original James-Stein paper (1960), which suggests that a priori interesting contrasts be separated out and estimated by maximum likelihood.

In a real data analysis, questions often arise which are neither prechosen nor purely a result of having looked at the data. As an example, consider the question of whether or not coffee drinking increases occurrence of pancreatic cancer. MacMahon et al. (1981) report an estimated relative risk for coffee drinkers that is significantly greater than 1, with significance in the range .01 through .05. The original subject of the MacMahon study was the relationship between smoking, alcohol consumption, and pancreatic cancer. The relationship with coffee consumption was, as the authors clearly state, unexpected. On the other hand it turns out that there is some older epidemiological evidence for a possible connection, and a single striking case history of husband-and-wife pancreatic cancer victims who had fortified their ground coffee with extra coffee syrup. Here we are in the uncomfortable middle ground between prechosen and purely post-hoc effects.

TABLE 2

Estimates of two differences between the baseball players. Difference 1 *is a prechosen contrast, while Difference* 2 *is chosen on the basis of the observed data. The estimate of standard error for the* MLE *is the usual one based on binomial variation. For the James-Stein method, the estimate of error is an empirical Bayes standard deviation, as explained in Efron and Morris* (1975).

	$\hat{\theta}_i$	$\tilde{\theta}_i$	θ_i
	MLE	JS	TRUE
Average, 9 Best Players	.289	.268	.293
Average, 9 Worst Players	.242	.258	.237
Difference 1	.047	.010	.056
(Error)	(.031)	(.016)	

	$\hat{\theta}_i$	$\tilde{\theta}_i$	θ_i
	MLE	JS	TRUE
Average, 9 Apparent Best	.321	.275	.270
Average, 9 Apparent Worst	.210	.251	.260
Difference 2	.111	.023	.010
(Error)	(.031)	(.016)	

The James-Stein estimate is a dramatic decision-theoretic contribution to estimation theory—as is the closely related theory of empirical Bayes estimation, Robbins (1956). It has not yet been incorporated into common statistical practice. Part of the delay relates to Remark *B*, the reluctance of scientists to report highly biased estimates. See Efron (1975a) and Efron and Morris (1972). A bigger question, in the author's opinion, relates to the pancreatic cancer example: how do we construct improved estimates in the "uncomfortable middle ground"? A successful answer is likely to be at least partly Bayesian while still enjoying good frequentist properties, as in the James-Stein estimate itself.

As a final example, consider fitting a polynomial regression model

$$x_i = \textstyle\sum_{j=0}^{J} \theta_j P_j(t_i) + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{8.4}$$

where the ε_i are i.i.d. $\mathcal{N}(0, \sigma^2)$, and $P_j(t)$ is a polynomial of degree j in the real variable t. The polynomials P_j are chosen to be orthogonal so that the least square (MLE) estimates $\hat{\theta}_j$ are mutually independent with equal variance. The value J is chosen deliberately large, so that we are certain that $\sum_{j=1}^{J} \theta_j P_j(t)$ is sufficient to describe the true regression function, but we suspect that a lower degree polynomial may be adequate.

Various hypothesis-testing sequences have been suggested to select the appropriate degree J_0 for the fitted polynomial $\sum_{j=0}^{J_0} \hat{\theta}_j P_j(t)$: step-down regression, step-up regression, etc.; see Chapter 6 of Draper and Smith (1981). These amount to specific recipes for going through the summary-comparison-inference cycle described in Section 3.

James-Stein estimation is a strong competitor in this situation, as are related methods such as ridge regression. We might estimate the first few coefficients, say θ_0, θ_1, θ_2, by maximum likelihood, and then shrink $\hat{\theta}_3, \cdots, \hat{\theta}_J$ towards 0, as in (8.3) but with $\hat{\theta}.$ replaced by 0. There are good theoretical grounds for believing that the regression estimated in this way is closer to the true regression than is that estimated by the hypothesis testing methods; see Sclove et al. (1972).

At this point we are verging on a new methodology for making summaries. The MLS is difficult to beat when the family of models $\mathscr{F}$ is well specified, but that is not the case for a stepwise fitted regression $\sum_{j=0}^{J_0} \hat{\theta}_j P_J(t)$. On the other hand, can we use the James-Stein estimated regression in a summary mode, reading off estimates of parameters of interest as we do with the MLS? Objections can be raised. For example, the estimate of the cubic component $\tilde{\theta}_3$ may be strongly biased toward zero. This can be unacceptable if it later turns out that the magnitude of the cubic effect is of crucial importance to the scientific interpretation of the experiment.

The James-Stein estimate suggests a more advanced approach to model building, and therefore to the summary of data using probability models, an approach which felicitously uses both Bayesian and frequentist ideas. So far this is just a suggestion. It will be no easy task to construct a comprehensive theory which incorporates both biased estimation and believable summary statistics.

9. Acknowledgement. Peter Bickel, Persi Diaconis, and David Hinkley have given me a great deal of useful advice in the preparation of this paper. Of course, they are not responsible for the way in which that advice was used.

REFERENCES

Barnard, G. (1974). Can we all agree on what we mean by estimation? *Utilitas Math.* **6** 3–22.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood! (with discussion). *Ann. Statist.* **8** 457–487.

Cox, D., and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *JRSS B* **39** 1–38.

Draper, N., and Smith, H. (1981). *Applied Regression Analysis*. Wiley, New York.

Efron, B. (1975a). Biased versus unbiased estimation. *Advances in Math.* **16** No. 3, 259–275. Reprinted in *Surveys in Applied Mathematics*, Academic, New York (1976).

Efron, B. (1975b). Defining the curvature of a statistical problem (with applications to second order efficiency), (with discussion). *Ann. Statist.* **3** 1189–1242.

Efron, B. (1978a). The geometry of exponential families. *Ann. Statist.* **6** 362–376.

Efron, B. (1978b). Regression and ANOVA with zero-one data: Measures of residual variation. *J. Amer. Statist. Assoc.* **73** 113–121.

Efron, B. (1978c). Controversies in the foundations of statistics. *Amer. Math. Monthly* **85** 241–246.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7,** No. 1, 1–26.

Efron, B. (1981a). Nonparametric estimates of standard error: The jackknife, the bootstrap, and other resampling methods. *Biometrika* 589–599.

Efron, B. (1981b). Nonparametric standard errors and confidence intervals. *Canadian Jour. Statist.* **8** 139–172.

Efron, B., and Hinkley, D. V. (1978). Assessing the accuracy of the MLE: Observed versus expected Fisher information (with discussion). *Biometrika* **65** 457–487.

Efron, B., and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimates—Part I: the Bayes case. *J. Amer. Statist. Assoc.* **66** 807–815.

Efron, B., and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimates—Part II: the empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139.

Efron, B., and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 177.

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Royal Soc. of London A* **222** 309–360.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Phil. Soc.* **22** 700–725.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Royal Soc. of London A* **144** 285–307.

Ghosh, J. K., and Sinha, B. K., and Wieand, H. S. (1980). Second order efficiency of the MLE with respect to any bounded bowl-shaped loss function. *Ann. Statist.* **8** 506–521.

Ghosh, J. K., and Subramanyan, K. (1974). Second order efficiency of maximum likelihood estimators. *Sankhya A* **36** 325–358.

Hinkley, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* **67** 287–292.

Hinkley, D. V. (1981). Likelihood. *Canad. J. Statist.* **8** 151–163.

Huber, P. (1981) *Robust Statistics*. Wiley, New York.

Jaeckel, L. (1972). The Infinitesimal Jackknife. Bell Labs. Memorandum #MM72-1215-11.

James, W., and Stein, C. (1960). Estimation with quadratic loss. *Proc. Fourth Berk. Symp.* **1** 361–379.

Jeffreys, H. (1967). *Theory of Probability*, 3rd ed. Clarendon, Oxford.

MacMahon, B., Yen, S. Trichopowlos, D., Warren, K., and Nordi, G. (1981). Coffee and cancer of the pancreas. *New England J. Med.* **304** 630–633.

Pfanzagl, J., and Wefelmeyer, W. (1978). A third order optimum property of the maximum likelihood estimator. *J. Mult. Analysis* **8** 1–29.

Rao, C. (1962). Efficient estimates and optimum inference problems in large samples (with discussion). *JRSS B* **24** 46–72.

Rao, C. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. 3rd Berk. Symp.* **1** 157–163.

SCLOVE, S. L., MORRIS, C., and RADHAKRISHNAN, R. (1972). Nonoptimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* **43** 1481–1490.

STEIN, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berk. Symp.* Vol. **I** 197–206. Berkeley, UC Press.

STIGLER, S. (1972). Do robust estimators work with *real* data? (with discussion). *Ann. Statist.* **6** 1055–1098.

TUKEY, J. (1977). *Exploratory Data Analysis.* Addison-Wesley, Indianapolis.

WOLFOWITZ, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28** 75–88.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CA 94305