The Annals of Statistics
1980, Vol. 8, No. 6, 1179–1197

THE IDENTIFICATION OF AN ELEMENT OF A LARGE POPULATION IN THE PRESENCE OF NOISE[1]

BY HERMAN CHERNOFF

Massachusetts Institute of Technology

A new approach is presented to the problem of determining whether an individual (the target) appears in a large file where individuals are identified by measurements subject to error. This approach attaches costs to searching and to missing the individual. It corresponds to testing a simple hypothesis, that the measurements on the target and an element in the library have a given joint distribution, against the alternative that they are independent. Certain measures of information from large deviation theory are relevant. There is a surprising reduction in effectiveness of information in the presence of error. Data compression issues are studied. Attention is paid to a two-stage search procedure where the file is subdivided into piles which are in turn subdivided into bins. Each pile is examined and either discarded or searched. If it is searched, each bin in it is examined and either discarded or searched. If a bin is searched, each element of the bin is compared with the target.

1. Introduction. Two identification or information retrieval problems of current interest are the following. Given the fingerprints of an individual, is this person already represented in a large government file of fingerprints? Given the mass spectrogram of a specimen of an organic chemical compound, which if any of the compounds listed in a library of mass spectra corresponds to the specimen? These questions raise serious difficulties when the data are subject to error.

We present an approach to these problems which relates to the theory of hypothesis testing and suggests the relevance of certain measures of information attached to the data. Let $X_i \in \mathcal{X}$, $i = 1, 2, \cdots, N$ be a large set of N points which corresponds to the population or *library* of individuals. We shall assume that these points are independent observations from a larger population with probability density $f_X(x)$. Let $Y \in \mathcal{Y}$ be a target point to be identified. That is, if X and Y are observations corresponding to the same individual, then (X, Y) has a joint distribution $f_{XY}(x, y)$ which indicates how X and Y are related. In the mass spectrogram problem X and Y may be two independently measured mass spectra for the same compound and except for measurement error, X and Y would tend to be very similar. Here $\mathcal{X}$ and $\mathcal{Y}$ would be identical spaces. Alternatively, the library might store abbreviated versions of the spectra in which case $\mathcal{X}$ and $\mathcal{Y}$ would be different but X and Y corresponding to the same compound would still be related.

The model proposed for our problem is the following. Given a target Y, we make a complete search of a region $\delta(Y) \subset \mathcal{X}$ at a cost of c per element $X_i \in \delta(Y)$. If the individual represented by the target is in the set $\{X_1, X_2, \cdots, X_N\}$ but not in $\delta(Y)$, a cost k is incurred for missing it. The prior probability that the individual is in the library $\{X_1, \cdots, X_N\}$ is π. Select the search region $\delta(Y)$ to minimize the expected cost.

This problem leads easily to a simple solution which identifies the problem with that of testing a simple hypothesis versus a simple alternative. However, our problem neglects some issues. It does not address directly the question of how to store the data in the computer to facilitate the search. Implicitly one assumes that there is no cost in identifying which elements

Received August 1978; revised July 1979.

[1] This paper was the subject of the Fisher Memorial Lecture presented to the ASA and IMS at the Atlanta, Georgia meeting in August 1975. This work was supported by the Office of Naval Research under contract N00014-75-C-0555(NR-0420331).

AMS 1970 *subject classifications.* Primary 68A50, 62F05; secondary 62E20, 62F20, 60F10.

Key words and phrases. Information retrieval, file, bin, library, large deviations, identification, information numbers, search.

X_i are in $\delta(Y)$ and should be compared with Y. One assumes the cost c pays for a definitive decision as to whether the item X_i corresponds to the same individual as Y. The possibility of stopping as soon as a definitive identification is made is ignored. Our analysis assumes that the individual appears at most once in the library. A few of these issues are attacked in later sections.

In Section 2, the optimal search region is described and related to the hypothesis testing problem. In Section 3 relevant bounds and information numbers are described and examples are presented in Section 4. These examples illustrate that a little "noise" or error in measurement has a surprisingly large effect in reducing the effective information available. Consequently a question of data compression arises, for it makes sense to save storage space by rounding off data that yields little effective information. Data compression is discussed in Section 5, where a conjecture is formulated which is related, in Section 6, to a variation of the classical isoperimetric problem.

The preceeding formulation assumes that there is no cost in determining whether a given X is in the set $\delta(Y)$ to be searched. This assumption is often inappropriate. In Section 7, we suggest dividing the library into subsets called *piles*. Then $\delta(Y)$ consists of a subset of piles each of which is either searched completely or not at all. Here the theory of the earlier sections applies directly and thus is relevant to computer file organization. A related question that arises in Section 8 is the construction of a natural metric on $\mathscr{X}$.

In Section 9, the decomposition of the library into piles is extended to a two-stage search procedure where the piles are further subdivided into *bins*. Some crude approximations are proposed and evaluated in some examples in Section 10. It is suggested that there is a premium on relatively noiseless and well-compressed data in the first stage.

Some of the work in fingerprint identification concentrates on how well a feature or observation divides up the population, and tends to ignore the effect of noise in reducing the effective information available. The examples in this paper point out that such noise is extremely influential. Hence the construction of a sound information retrieval scheme requires the explicit calculation of the effects of noise.

A discussion of some applications of the results in this paper is presented in [5].

2. The optimal search region. In this section we determine the optimal search region for the simple problem formulated in the introduction. We relate the problem and its costs to those of testing the simple hypothesis, that two random variables have a given joint distribution, vs. the simple alternative, that they are independent with the corresponding marginal distributions.

For simplicity we shall engage in a slight abuse of notation where X and Y are treated as continuous random variables with marginal densities $f_X(x)$ and $f_Y(y)$ and joint density $f_{XY}(x, y)$ with respect to Lebesgue measure. Our treatment also applies to more general random variables.

Given the cost structure described in the introduction, the cost associated with a search region defined by the function δ is

$$C = \int C(y) f_Y(y)\, dy \tag{2.1}$$

where $C(y)$, the conditional expected cost given $Y = y$, is

$$C(y) = cNP\{X \in \delta(y)\} + k\pi P\{X \notin \delta(y) \mid Y = y\}. \tag{2.2}$$

We see that

$$\begin{aligned} C(y) &= cN \int_{\delta(y)} f_X(x)\, dx + k\pi \int_{\delta^c(y)} \frac{f_{XY}(x, y)}{f_Y(y)}\, dx \\ &= k\pi \left(1 + \int_{\delta(y)} \left[\lambda_0 f_X(x) - \frac{f_{XY}(x, y)}{f_Y(y)}\right] dx\right) \end{aligned}$$

where

$$\lambda_o = cN/k\pi. \tag{2.3}$$

It is clear that an optimal search region is defined by

$$\delta(y) = \{x : \lambda(x, y) \geq \lambda_o\} \tag{2.4}$$

where

$$\lambda(x, y) = \frac{f_{XY}(x, y)}{f_X(x)f_Y(y)}. \tag{2.5}$$

Furthermore the associated expected cost is

$$C = k\pi \left(\iint_{\lambda \geq \lambda_o} \lambda_o f_X(x) f_Y(y)\, dx\, dy + \iint_{\lambda < \lambda_o} f_{XY}(x, y)\, dx\, dy \right)$$
$$C = k\pi[\alpha + \lambda_o \beta] \tag{2.6}$$

where

$$\alpha = P_1\{\lambda(X, Y) < \lambda_o\}, \tag{2.7}$$

$$\beta = P_2\{\lambda(X, Y) \geq \lambda_o\}, \tag{2.8}$$

P_1 is the probability measure corresponding to the density $f_{XY}(x, y)$ and P_2 is that corresponding to $f_X(x)f_Y(y)$.

In summary, our optimal choice of search region corresponds to a likelihood-ratio test of the simple hypothesis

$$H_1 : (X, Y) \sim f_{XY}(x, y)$$

versus the simple alternative

$$H_2 : (X, Y) \sim f_X(x)f_Y(y)$$

and the expected cost is a linear function of the error probabilities α and β.

In retrospect it is not surprising that we should choose to confine our search to that region of $\mathscr{X}$ for which the "observations", X, Y would lead to accepting the hypothesis H_1 that (X, Y) are from the joint distribution corresponding to the same individual versus the alternative H_2 that X and Y are independent with the specified marginal distributions.

Similar search strategies have occasionally been suggested in the past, usually with little or no rationale. One exception is that due to Sunter and Felleghi [8] where a slightly more elaborate but related approach is made in connection with the problem of record linkage.

One important characteristic of our problem is that in typical examples N is very large, with the consequence that λ_o is also very large. Thus, in the hypothesis testing context, acceptance of H_1 occurs only when one is very sure that (X, Y) are not independent. In this way the set $\delta(Y)$ is a relatively *small* set of points X very "close" to Y, in terms of the measure λ, so that our search does not have to cover too many elements in the library.

3. Bounds and information numbers. In this section, a bound on $\alpha + \lambda_o\beta$ is derived. Its relation to large deviation theory and certain measures of information is discussed. These information numbers permit one to assess how useful certain components of the vectors X and Y are in reducing the cost C.

We have noted in (2.6) that $C = k\pi(\alpha + \lambda_o\beta)$. It is easy to see that for the optimal δ

$$\alpha + \lambda_o\beta = \iint \min[\lambda_o f_X(x)f_Y(y), f_{XY}(x, y)]\, dx\, dy.$$

Since $\min(a, b) \leq a^t b^{1-t}$ for $a > 0$, $b \geq 0$, and $0 \leq t \leq 1$,

$$\alpha + \lambda_o\beta \leq \inf_{0\leq t\leq 1}\lambda_o^t \int\int f_{XY}^{1-t}(x, y)f_X^t(x)f_Y^t(y)\,dx\,dy. \tag{3.1}$$

A weaker bound of interest is that where the infimum on the right is replaced by the value with $t = \frac{1}{2}$.

Inequality (3.1) is more than a mere upper bound on $\alpha + \lambda_o\beta$. To appreciate it better, let us consider some results of large deviation theory applied to testing a simple hypothesis H_1: $f = f_1$ versus a simple alternative $H_2: f = f_2$ based on a large number n of i.i.d. observations on a random variable U with distribution f. For simplicity we shall assume that these two distributions are absolutely continuous with respect to one another, i.e., that $\int_{f_2=0} f_1(x)\,dx = \int_{f_1=0} f_2(x)\,dx = 0$. For each likelihood-ratio test there is a corresponding (α, β) and it is well known [3, 4] that

$$-\frac{1}{n}\inf[\log(\alpha + \beta)] \to I_o^*(f_1, f_2) \tag{3.2}$$

where

$$I_o^*(f_1, f_2) = -\log\left[\inf_{0\leq t\leq 1}\int f_1^{1-t}(x)f_2^t(x)\,dx\right] \tag{3.2}'$$

and that the infimum of the integral is attained for $0 \leq t \leq 1$. This means that as n becomes large, $\alpha + \beta$ approaches zero exponentially fast at a rate determined by I_o^*.

Moreover if we keep α fixed at α_o

$$-\frac{1}{n}\inf_{\alpha=\alpha_o}[\log\beta] \to I_K^*(f_1, f_2) \tag{3.3}$$

where

$$I_K^*(f_1, f_2) = \int \log\left(\frac{f_1(x)}{f_2(x)}\right)f_1(x)\,dx. \tag{3.3}'$$

Here I_K^* is the *Kullback-Leibler information* number for discriminating between f_1 and f_2 when H_1 is true and I_o^* is sometimes referred to as the *Chernoff information*. Incidentally, when the infimum of the integral in (3.2)′ is replaced by the value for $t = \frac{1}{2}$, the negative of the resulting integral is $I_B^*(f_1, f_2)$ which is associated with the *Bhattacharya distance* and the *Matusita distance* or *affinity* [1, 7].

Suppose now that λ_o increases exponentially with n, so that

$$\frac{1}{n}\log\lambda_o = \gamma_o. \tag{3.4}$$

It can be shown that if $I_K^*(f_1, f_2) > \gamma_o \geq 0$, it is possible to have α and $\beta e^{n\gamma_o}$ both approach zero. The argument deriving (3.2) is easily extended to give

$$-\frac{1}{n}\inf[\log(\alpha + e^{n\gamma_o}\beta)] \to I_{\gamma_o}^*(f_1, f_2) \tag{3.5}$$

where

$$I_{\gamma_o}^*(f_1, f_2) = -\log\left[\inf_{0\leq t\leq 1}\int f_1^{1-t}(x)f_2^t(x)e^{t\gamma_o}\,dx\right]. \tag{3.5}'$$

The integral in (3.5)′ attains its minimum for $0 \leq t \leq 1$ because the integral is convex, assumes the values 1 and $\exp(\gamma_o) \geq 1$ *for* $t = 0$ and 1 respectively, and its derivative at $t = 0$ is $\gamma_o - I_K^*(f_1, f_2) < 0$. The information number $I_{\gamma_o}^*$ measures the exponential rate at which $\alpha + \lambda_o\beta$ approaches zero when $\lambda_o \approx \exp(n\gamma_o)$.

These properties of the information numbers I_k^*, I_o^*, $I_{\gamma_o}^*$ and I_B^* may be derived with reasonable ease using the basic large deviation result [3] that if $a \leq E(Z)$

$$-\frac{1}{n}\log[P(\bar{Z} \geq a)] \to -\log[\inf_{t\geq 0} E\{e^{t(Z-a)}\}] = \rho(a) \tag{3.6}$$

where $\bar{Z}$ is the average of n i.i.d. observations on Z. This result is applied to $Z = \log[f_1(X)/f_2(X)]$.

Several relations among these information numbers follow immediately from the above results. For example,

$$\begin{gathered} I_K^* \geq I_o^* \geq I_{\gamma_o}^* \geq \lambda_o^{1/2} I_B^*. \\ I_o^* \geq I_B^*. \end{gathered} \tag{3.7}$$

If we think of X and Y in our information retrieval problem as vectors, $X = (U_1, U_2 \cdots, U_n)$ and $Y = (V_1, V_2, \cdots, V_n)$ where the (U_i, V_i) are i.i.d. random variables with density f_{UV}, then

$$I_K^*(f_{UV}, f_U f_V) = I_K(f_{UV}) \tag{3.8}$$

where

$$I_K(f_{UV}) := \int f_{UV}\log(f_{UV})\, du\, dv - \int f_U \log(f_U)\, du - \int f_V \log(f_V)\, dv \tag{3.8}'$$

is the Shannon mutual information. Moreover

$$I_{\gamma_o}^*(f_{UV}, f_U f_V) = I_{\gamma_o}(f_{UV}) \tag{3.9}$$

where

$$I_{\gamma_o}(f_{UV}) := -\log\left[\inf_{0\leq t\leq 1}\lambda_o^t \int f_{UV}^{1-t} f_U^t f_V^t\, du\, dv\right], \tag{3.9}'$$

and

$$I_B^*(f_{UV}, f_U f_V) = I_B(f_{UV}) \tag{3.10}$$

and

$$I_B(f_{UV}) = -\log\left[\int f_{UV}^{1/2} f_U^{1/2} f_V^{1/2}\, du\, dv\right]. \tag{3.10}'$$

Finally equation (3.7) reduces to

$$\begin{gathered} I_K \geq I_o \geq I_{\gamma_o} \geq \lambda_o^{1/2} I_B \\ I_o \geq I_B. \end{gathered} \tag{3.11}$$

Thus, the right-hand side of (3.1) is a coarse approximation to, as well as an upper bound for $\alpha + \lambda_o\beta$, at least in the case where (X, Y) is a vector of i.i.d. random vectors (U_i, V_i). Furthermore I_{γ_o} measures the *effective* information in (U_i, V_i) in the sense that the cost C is reduced roughly by $\exp(-I_{\gamma_o})$ when (U_i, V_i) are available in the identification problem.

This approximation is coarse in the sense that the ratio of the left side to the right of (3.1) could be a power of n. More specifically, in the context of equation (3.6) Cramér [6] has shown that for absolutely continuous distributions (3.6) may be refined to

$$P(\bar{Z} \geq a) \approx \frac{1}{n^{1/2} t^*(a)} e^{-n\rho(a)} \tag{3.12}$$

where $t^*(a)$ is the minimizing value of t in (3.6). Under more general conditions [2] we know that $P(\bar{Z} \geq a) \sim n^{-1/2}\exp(-n\rho(a))$. Thus one may have some insight on the extent to which the right-hand side of (3.1) overestimates the left. For integer valued random variables, Blackwell and Hodges [2] derived a result different from but similar to (3.12).

4. Information numbers in some examples. We consider the relevant information numbers for a few simple examples. The accompanying tables illustrate two main related points. A little

noise degrades the information content to an extent which the author finds surprising. Also, when noise is present the Shannon information I_K generally tends to be much larger than the more relevant information numbers I_{γ_o} and I_B.

EXAMPLE 4.1. Suppose that $P(Z = 1) = p$ and $P(Z = 0) = 1 - p$ and X and Y are independent observations on Z which have error probabilities ϵ_o and ϵ_1 when $Z = 0$ and $Z = 1$ respectively. That is, $P(X = 1) = P(Y = 1) = p(1 - \epsilon_1) + (1 - p)\epsilon_o$, $P(X = Y = 1) = p(1 - \epsilon_1)^2 + (1 - p)\epsilon_o^2$, $P(X = 1, Y = 0) = p(1 - \epsilon_1)\epsilon_1 + (1 - p)\epsilon_o(1 - \epsilon_o)$, etc. In Table 4.1, we present values of I_K, I_o, I_B, $I_{0.1}$ and $I_{0.2}$ for a few cases of $(p, \epsilon_1, \epsilon_o)$. The corresponding values of t, t_o, $t_{0.1}$ and $t_{0.2}$, at which the minima in (3.1) are attained are also presented.

If $p = 0.5$ and ϵ_o and ϵ_1 approach 0, $I_K \to \log 2$, i.e., one bit. It also can be shown that $I_o \to \log 2$ and $I_B \to \frac{1}{2} \log 2$. Nevertheless, even when ϵ_o and ϵ_1 are both as small as 0.02, I_o is much closer to I_B than to I_K. Then $I_o = 0.187$ which is only the equivalent of about a quarter of a bit. If $\epsilon_o = \epsilon_1 = 0.1$ then I_o is only 0.062 which is only one third as much.

Note that I_B is often close to I_o, and $I_{\gamma_o} + \gamma_o/2$ exceeds, but is often reasonably close to, I_B especially when t_{γ_o} is close to 0.5. Note also that when ϵ_o and ϵ_1 approach zero, then for general p, $I_K \to -[p \log p + (1 - p)\log(1 - p)]$, $I_o \to -\log[p^2 + (1 - p)^2]$ and $I_B \to -\log[p^{3/2} + (1 - p)^{3/2}]$.

TABLE 4.1
Information numbers for Example 4.1

$$P(Z = 1) = p \quad P(Z = 0) = 1 - p$$

$$P(X = 0 \mid Z = 1) = P(Y = 0 \mid Z = 1) = \epsilon_1$$

$$P(X = 1 \mid Z = 0) = P(Y = 1 \mid Z = 0) = \epsilon_o$$

p	ϵ_1	ϵ_o	I_K	I_o	I_B	$I_{0.1}$	$I_{0.2}$	t_o	$t_{0.1}$	$t_{0.2}$
0.05	0.002	0.002	0.181	0.048	0.047	0.009	0.000[1]	0.56	0.22	0.00
		0.020	0.119	0.026	0.026	0.001	0.000	0.48	0.06	0.00
		0.100	0.038	0.009	0.009	0.000	0.000	0.48	0.00	0.00
	0.020	0.002	0.170	0.043	0.043	0.006	0.000	0.54	0.19	0.00
		0.020	0.112	0.025	0.025	0.003	0.000	0.48	0.04	0.00
		0.100	0.035	0.008	0.008	0.000	0.000	0.48	0.00	0.00
	0.100	0.002	0.132	0.030	0.030	0.001	0.000	0.50	0.09	0.00
		0.020	0.088	0.019	0.019	0.000	0.000	0.47	0.00	0.00
		0.100	0.025	0.006	0.006	0.000	0.000	0.48	0.00	0.00
0.20	0.002	0.002	0.477	0.187	0.176	0.128	0.076	0.62	0.56	0.47
		0.020	0.386	0.120	0.119	0.070	0.031	0.56	0.45	0.32
		0.100	0.198	0.053	0.053	0.013	0.000	0.51	0.27	0.00
	0.020	0.002	0.439	0.153	0.149	0.098	0.053	0.59	0.51	0.40
		0.020	0.360	0.108	0.107	0.059	0.024	0.55	0.43	0.28
		0.100	0.184	0.049	0.049	0.010	0.000	0.51	0.25	0.00
	0.100	0.002	0.319	0.092	0.092	0.045	0.014	0.54	0.40	0.23
		0.020	0.265	0.072	0.072	0.029	0.005	0.52	0.34	0.14
		0.100	0.131	0.034	0.034	0.002	0.000	0.50	0.12	0.00
0.50	0.002	0.002	0.667	0.317	0.287	0.230	0.172	0.63	0.59	0.55
		0.020	0.589	0.228	0.219	0.171	0.119	0.60	0.54	0.49
		0.100	0.382	0.118	0.117	0.069	0.031	0.54	0.44	0.31
	0.020	0.002	0.589	0.228	0.219	0.171	0.119	0.60	0.54	0.49
		0.020	0.528	0.187	0.183	0.133	0.085	0.57	0.51	0.44
		0.100	0.345	0.103	0.103	0.056	0.021	0.54	0.42	0.27
	0.100	0.002	0.382	0.118	0.117	0.069	0.031	0.54	0.44	0.31
		0.020	0.345	0.103	0.103	0.056	0.021	0.54	0.42	0.27
		0.100	0.222	0.062	0.062	0.020	0.001	0.52	0.31	0.06

[1] When $\gamma \geq I_K$, $I_\gamma = 0$ and $t_\gamma = 0$.

EXAMPLE 4.2. Related to Example 4.1 is the case where X and Y can be 0 or 1 and we simply present $P(X = 1) = P(Y = 1) = p$ and $P(X = Y = 1) = p - \epsilon$. Then the other joint probabilities are $P(X = 1, Y = 0) = P(X = 0, Y = 1) = \epsilon$, and $P(X = Y = 0) = 1 - p - \epsilon$. In Table 4.2 we present values I_K, I_o, I_B, $I_{0.1}$, $I_{0.2}$, t_o, $t_{0.1}$ and $t_{0.2}$ for a few cases of (p, ϵ).

EXAMPLE 4.3. Let Z be normally distributed with mean 0 and variance 1. Let X and Y represent independent observations on Z with normal errors. That is $X = Z + W_1$ and $Y = Z + W_2$ where Z, W_1, and W_2 are independent and normal with mean 0 and W_1 and W_2 have variances σ_1^2 and σ_2^2.

Table 4.3 presents values of I_K, I_o, I_B, $I_{0.1}$, $I_{0.2}$, t_o, $t_{0.1}$ and $t_{0.2}$ for several values of σ_1 and σ_2.

In this example I_K and I_B may be computed to be

$$I_K = \frac{1}{2} \log \left(\frac{1 + d}{d} \right) \tag{4.1}$$

and

$$I_B = \frac{1}{4} \log \left[\frac{(0.75 + d)^2}{d(1 + d)} \right] \tag{4.2}$$

where

$$d = \sigma_1^2 + \sigma_2^2 + \sigma_1^2 \sigma_2^2. \tag{4.3}$$

For $d \to 0$, $I_o \approx I_K \approx -\frac{1}{2} \log d$. However, as in Example 4.1, I_o is much closer to I_B than to I_K when σ_1 and σ_2 are as small as 0.02 ($d = 0.0008$). When $\sigma_1 = \sigma_2 = 0.1$, I_K/I_o is almost 2 and I_o is approximately the equivalent of only 1½ bits.

5. Data compression. Since the amount of effective information seems remarkably small in some of the examples considered, a problem that arises naturally is that of data compression. In Example 4.3, if $\sigma_1 = \sigma_2 = 0.1$, I_o is the equivalent of 1½ bits. Why should one store X and Y to several significant figures using many bits of storage space, when the amount of effective

TABLE 4.2

Information numbers for Example 4.2

$$P(X = Y = 1) = p - \epsilon$$

$$P(X = 1, Y = 0) = P(X = 0, Y = 1) = \epsilon$$

$$P(X = Y = 0) = 1 - p - \epsilon$$

p	ϵ	I_K	I_o	I_B	$I_{0.1}$	$I_{0.2}$	t_o	$t_{0.1}$	$t_{0.2}$
0.02	0.01	0.028	0.005	0.005	0.000	0.000	0.42	0.00	0.00
0.05	0.01	0.118	0.026	0.026	0.001	0.000	0.49	0.06	0.00
	0.02	0.068	0.014	0.014	0.000	0.000	0.46	0.00	0.00
0.10	0.01	0.238	0.063	0.063	0.021	0.002	0.53	0.31	0.08
	0.02	0.179	0.044	0.044	0.008	0.000	0.50	0.21	0.00
	0.05	0.063	0.014	0.014	0.000	0.000	0.47	0.00	0.00
0.30	0.01	0.515	0.187	0.181	0.131	0.083	0.59	0.52	0.45
	0.02	0.447	0.147	0.144	0.095	0.052	0.56	0.48	0.38
	0.05	0.296	0.084	0.084	0.039	0.010	0.53	0.38	0.20
	0.10	0.133	0.034	0.034	0.002	0.000	0.51	0.13	0.00
	0.20	0.001	0.000	0.000	0.000	0.000	0.50	0.00	0.00
0.50	0.01	0.595	0.233	0.223	0.175	0.123	0.60	0.55	0.49
	0.02	0.525	0.186	0.181	0.131	0.084	0.57	0.51	0.43
	0.05	0.368	0.112	0.112	0.064	0.027	0.11	0.06	0.03
	0.10	0.193	0.053	0.053	0.013	0.000	0.05	0.01	0.00
	0.20	0.020	0.005	0.005	0.000	0.000	0.50	0.00	0.00

TABLE 4.3
Information numbers for Example 4.3

$X = Z + W_1$ $\quad Y = Z + W_2$
$\mathcal{L}(Z) = N(0, 1)$ $\quad \mathcal{L}(W_i) = N(0, \sigma_i^2)$

σ_1	σ_2	I_K	I_o	I_B	$I_{0.1}$	$I_{0.2}$	t_o	$t_{0.1}$	$t_{0.2}$
0.02	0.02	3.566	2.320	1.639	2.230	2.139	0.93	0.93	0.93
0.02	0.05	2.923	1.845	1.318	1.761	1.677	0.86	0.86	0.85
0.02	0.10	2.288	1.329	1.002	1.249	1.169	0.81	0.80	0.79
0.02	0.20	1.624	0.829	0.674	0.754	0.681	0.76	0.74	0.73
0.02	0.40	0.989	0.419	0.373	0.352	0.288	0.68	0.65	0.62
0.05	0.05	2.651	1.622	1.183	1.539	1.457	0.84	0.83	0.82
0.05	0.10	2.196	1.257	0.956	1.177	1.098	0.80	0.80	0.79
0.05	0.20	1.599	0.811	0.662	0.737	0.663	0.76	0.74	0.72
0.05	0.40	0.933	0.415	0.370	0.348	0.285	0.68	0.65	0.62
0.10	0.10	1.964	1.079	0.841	1.000	0.923	0.79	0.78	0.77
0.10	0.20	1.518	0.755	0.623	0.682	0.609	0.75	0.73	0.71
0.10	0.40	0.960	0.402	0.360	0.336	0.273	0.68	0.65	0.61
0.20	0.20	1.292	0.604	0.515	0.533	0.464	0.72	0.70	0.68
0.20	0.40	0.883	0.359	0.325	0.294	0.233	0.67	0.63	0.59
0.40	0.40	0.680	0.253	0.237	0.192	0.137	0.58	0.52	0.44

information is so little? Suppose one used only one or two bits of storage space to store part of the data in X and Y. Would I_o be reduced very much then?

Let us elaborate on Example 4.3.

EXAMPLE 5.1. For Example 4.3 with $\sigma_1 = \sigma_2 = \sigma$ select $k - 1$ real numbers in increasing order $a_1 < a_2 < \cdots < a_{k-1}$. Let $a_o = -\infty$ and $a_k = \infty$ and let $X^* = i$ if $a_{i-1} < X(1 + \sigma^2)^{-1/2} \leq a_i$ and let $Y^* = i$ if $a_{i-1} < Y(1 + \sigma^2)^{-1/2} \leq a_i$. We select the a_i so that the intervals maximize the various information numbers subject to the symmetry constraint $a_i + a_{k-i} = 0$. In Table 5.1 we list the optimal I_K, I_o, I_B, $I_{0.1}$ and $I_{0.2}$. This is repeated for several values of σ and k. The end points of the optimal symmetric intervals of $X(1 + \sigma^2)^{-1/2}$ have been computed for I_K, I_o, I_B, $I_{0.1}$, and $I_{0.2}$ for various values of k and σ. The optimal intervals for I_B, $I_{0.1}$ and $I_{0.2}$ are generally very close to those for I_o. The information numbers are relatively insensitive to changes in the intervals. For σ as large as 0.2, the optimal intervals for I_o put approximately equal X probability in each interval. For σ as small as 0.05 the optimal end intervals for I_o tend to have more than twice as much probability as the others each of which are roughly equal.

Another variation of Example 5.1 is that where the data in the library are compressed but Y is not. In this variation we would select the intervals to maximize $I_o(f_{X \cdot Y})$, $I_K(f_{X \cdot Y})$, etc. Results for this variation appear in Table 5.2.

From Table 5.1 one observes that when $\sigma_1 = \sigma_2 = 0.1$, I_o for one bit of storage space ($k = 2$), is less than half of I_o for two bits of storage space ($k = 4$). Consequently we have the following peculiar situation. If (Z_1, Z_2) were i.i.d. $N(0, 1)$ random variables and (X_1, X_2) and (Y_1, Y_2) were independent observations on (Z_1, Z_2) with measurement standard deviation 0.1, then using one bit of storage space on each of X_1, X_2, Y_1 and Y_2 is less effective than using two bits on each of X_1 and Y_1 and *discarding* X_2 and Y_2.

In the chemical identification problem the mass spectrogram is ordinarily a vector of about 200 to 400 components. If we wish to store fewer than 100 bits of information it would seem important to answer the following question related to the above mentioned phenomenon. Suppose Z_1 and Z_2 are independent $N(0, 1)$ variables and $X_i = Z_i + W_{i1}$ and $Y_i = Z_i + W_{i2}$ where the W_{ij} are independent $N(0, \sigma^2)$ random variables, $i = 1, 2, j = 1, 2$. It is desired to summarize (X_1, X_2) and (Y_1, Y_2), with one bit of data for each pair. How can this be most informatively arranged? To be more specific, let R be a region in E_2 and let $X^* = 1$ if $(X_1, X_2) \in R$ and 0 otherwise. Let $Y^* = 1$ if $(Y_1, Y_2) \in R$ and 0 otherwise. Select R to maximize

TABLE 5.1
Information numbers using k intervals in Example 5.1

$X = Z + W_1$ $\quad Y = Z + W_2$
$\mathcal{L}(Z) = N(0, 1)$ $\quad \mathcal{L}(W_i) = N(0, \sigma^2)$

k	σ	I_K	I_o	I_B	$I_{0.1}$	$I_{0.2}$	t_o	$t_{0.1}$	$t_{0.2}$
2	0.05	0.586	0.225	0.217	0.168	0.117	0.59	0.54	0.48
	0.10	0.510	0.177	0.173	0.123	0.077	0.57	0.50	0.42
	0.20	0.394	0.123	0.122	0.073	0.035	0.55	0.44	0.33
4	0.05	1.142	0.593	0.460	0.510	0.428	0.90	0.88	0.86
	0.10	0.983	0.487	0.394	0.411	0.338	0.77	0.74	0.71
	0.20	0.758	0.328	0.262	0.262	0.200	0.68	0.64	0.59
6	0.05	1.444	0.840	0.606	0.752	0.654	0.90	0.89	0.83
	0.10	1.232	0.659	0.520	0.581	0.506	0.78	0.76	0.74
	0.20	0.940	0.429	0.374	0.360	0.295	0.71	0.68	0.64
8	0.05	1.643	1.028	0.707	0.941	0.855	0.92	0.91	0.90
	0.10	1.394	0.776	0.600	0.694	0.616	0.81	0.80	0.77
	0.20	1.044	0.485	0.418	0.415	0.348	0.72	0.69	0.66
9	0.05	1.721	1.086	0.748	0.997	0.912	0.91	0.89	0.89
	0.10	1.456	0.822	0.630	0.734	0.657	0.82	0.80	0.78
	0.20	1.080	0.504	0.432	0.434	0.366	0.73	0.70	0.67
∞	0.05	2.651	1.622	1.183	1.539	1.457	0.84	0.83	0.82
	0.10	1.964	1.079	0.841	1.000	0.923	0.79	0.78	0.77
	0.20	1.292	0.604	0.515	0.533	0.464	0.72	0.70	0.68

$I_o(f_{X^*Y^*})$ or $I_K(f_{X^*Y^*})$. If the following conjecture is true, optimality can be achieved by discarding X_2 and Y_2.

CONJECTURE. An optimal R consists of $\{(x_1, x_2): x_1 \leq 0\}$.

This conjecture seems rather paradoxical. For a classical statistician it is difficult to accept the statement that no gain can be derived from using the information available in X_2 even when one is confined to using only one bit of storage space. A precise proof of this conjecture does not exist but partial results described in the next section tend to support it.

In the meantime some doubt may be cast on the intuition of the statistician who would suggest that a better alternative candidate for an optimal R would be $R_1 = \{(x_1, x_2): x_1 + x_2 \leq 0\}$. Then $\mathcal{L}(X_1 + X_2, Y_1 + Y_2) = \mathcal{L}(2^{1/2}(X_1, Y_1))$ and hence, for any method depending on $X_1 + X_2$ there is an equally informative one using X_1. Thus if R_1 is optimal so is $R_2 = \{(x_1, x_2): x_1 \leq 0\}$.

We conclude this section with the bare outline of an example of data compression.

EXAMPLE 5.2. Let Z_i, U_i, V_i, $i = 1, \cdots, n$ be i.i.d. $N(0, 1)$ random variables. Let $X_i = Z_i + \sigma U_i$ and $Y_i = Z_i + \sigma V_i$ for $1 \leq i < n$. Let X^* and Y^* be the indices of the largest X_i and Y_i respectively.

Then the information in using the (X, Y) vectors is n times that of a single (X_i, Y_i). Using X^* and Y^* involves only $\log_2 n$ bits of storage space. If σ is very small the effective information (I_o or I_K) is close to $\log n$, which is equivalent to $\log_2 n$ bits. Thus while X^* and Y^* present only a small part of the available information, they present that part efficiently with respect to the use of storage space.

6. Comments on the conjecture of Section 5. We have remarked on the paradoxical nature

TABLE 5.2

Information numbers using k intervals in variation of Example 5.1 where Y is not compressed

$X = Z + W_1 \qquad Y = Z + W_2$

$\mathcal{L}(Z) = N(0, 1) \qquad \mathcal{L}(W_i) = N(0, \sigma^2)$

k	σ	I_K	I_o	I_B	$I_{0.1}$	$I_{0.2}$	t_o	$t_{0.1}$	$t_{0.2}$
2	0.05	0.668	0.465	0.308	0.378	0.293	0.88	0.86	0.83
	0.10	0.643	0.363	0.271	0.284	0.208	0.80	0.77	0.73
	0.20	0.595	0.243	0.207	0.174	0.111	0.71	0.66	0.59
4	0.05	1.255	0.881	0.596	0.794	0.708	0.87	0.86	0.85
	0.10	1.128	0.670	0.509	0.591	0.513	0.80	0.78	0.76
	0.20	0.893	0.428	0.366	0.358	0.291	0.71	0.68	0.65
6	0.05	1.587	1.082	0.747	0.996	0.912	0.86	0.85	0.85
	0.10	1.392	0.809	0.620	0.730	0.653	0.79	0.78	0.77
	0.20	1.054	0.500	0.428	0.429	0.362	0.72	0.69	0.66
8	0.05	1.804	1.194	0.843	1.110	1.033	0.86	0.85	0.84
	0.10	1.548	0.886	0.684	0.807	0.730	0.79	0.78	0.77
	0.20	1.135	0.535	0.458	0.464	0.396	0.72	0.69	0.67
10	0.05	1.958	1.290	0.909	1.206	1.122	0.85	0.84	0.84
	0.10	1.648	0.934	0.724	0.856	0.779	0.79	0.78	0.77
	0.20	1.181	0.555	0.475	0.484	0.416	0.72	0.70	0.67
∞	0.05	2.651	1.622	1.183	1.539	1.457	0.84	0.83	0.82
	0.10	1.964	1.079	0.841	1.000	0.923	0.79	0.78	0.77
	0.20	1.292	0.604	0.515	0.533	0.464	0.72	0.70	0.68

of the conjecture of Section 5. We shall outline some of the ideas that appear in a partial proof of that conjecture leaving the details for subsequent publication.

Let $X^* = 1$ if $(X_1, X_2) \in R$ and 0 otherwise and let $Y^* = 1$ if $(X_1, X_2) \in R$ and 0 otherwise. If σ is small, the information I_o is log 2 minus an increment, which depends on $\epsilon = P(X^* = 1, Y^* = 0)$ and $\eta = P(X^* = 1) - \frac{1}{2}$, and which is small when ϵ and η are small. Ideally one would like to have both ϵ and η be zero but ϵ cannot go below some number depending on σ. For fixed η, it is preferable to minimize ϵ which is approximately proportional to the integrated density along the boundary of R. Thus a proof of the conjecture is related to the demonstration that the solution of the following variational problem is given by $R = R(\eta) =$ a half plane.

PROBLEM. Find the region R with boundary B which minimizes

$$L = \int_B \phi(x_1)\phi(x_2)(dx_1^2 + dx_2^2)^{1/2}$$

subject to

$$A = \int_R \phi(x_1)\phi(x_2)\, dx_1\, dx_2 = \tfrac{1}{2} + \eta,$$

where $\phi(x)$ is the standard normal density $(2\pi)^{-1/2}\exp(-x^2/2)$.

Note that when $\phi(x)$ is replaced by 1, L and A are the perimeter and area of R and we have the classical isoperimetric problem. A proof that the half plane is a stationary solution of our problem involves the following lemma which is of interest in its own right and may be proved by using a Hermite expansion.

LEMMA. *If Z is normally distributed with mean 0 and variance 1, and if g is an absolutely differentiable function such that $g(Z)$ has finite variance, then*

$$E[g'(Z)]^2 \geq \mathrm{Var}[g(Z)]$$

with equality if and only if g is linear.

7. Piles. Up to this point, the major issue in the case of very large libraries has been avoided. In that case it is not generally feasible to decide in advance whether an element X of $\mathscr{X}$ lies in $\delta(y)$ or not. To implement our procedure involves the possibly prohibitive additional cost of c_oN where c_o is the average cost of determining whether $X \in \delta(Y)$. Nevertheless, this issue may be confronted without abandoning the theory discussed before. The following considerations are relevant to the problem of file storage.

Let us suppose that $\mathscr{X}$ is partitioned into disjoint subsets $\mathscr{X}_1, \mathscr{X}_2, \cdots, \mathscr{X}_M$ with $N_1, N_2, \cdots, N_M$ elements respectively. We shall call these subsets *piles*. When $Y = y$ is observed, a decision is made for each pile $\mathscr{X}_i$ as to whether it should be skipped or searched completely at a cost of c per element. Then $\delta(y)$ consists of a union of some of the $\mathscr{X}_i$ and the cost c_oN of the previous paragraph is replaced by c_o^*M, where c_o^* is the cost of making the decision for each $\mathscr{X}_i$. For example, if the size of each $\mathscr{X}_i$ is of the order of magnitude of 10^3 and c_o^* is comparable to c_o then M is of the order of $10^{-3}N$ and the cost c_oN is reduced by a factor of 10^{-3}.

Two questions must be answered. How is it to be determined whether $\mathscr{X}_i \subset \delta(y)$? How does this procedure affect the cost $C(y)$ and C discussed in Section 2?

As in Section 2, it is clear that

$$C = \pi k\left[\int_{x\in\delta(y)} f_{XY}(x, y)\,dx\,dy + \lambda_o \int_{x\notin\delta(y)} f_X f_Y\,dx\,dy\right].$$

If we let $X^* = i$ if $X \in \mathscr{X}_i$, and define $\delta^*(y) = \{i: \mathscr{X}_i \subset \delta(y)\}$

$$f_{X^*Y}(i, y) = \int_{x\in\mathscr{X}_i} f_{XY}(x, y)\,dx$$

$$f_{X^*}(i) = \int_{x\in\mathscr{X}_i} f_X(x)\,dx.$$

Then

$$C = \pi k \int [\textstyle\sum_{i\in\delta^*(y)} f_{X^*Y}(i, y) + \lambda_o \sum_{i\notin\delta^*(y)} f_{X^*}(i) f_Y(y)]\,dy$$

which attains its minimum value of

$$C^* = \pi k \int \textstyle\sum_i \min[f_{X^*Y}(i, y), \lambda_o f_{X^*}(i) f_Y(y)]\,dy \tag{7.1}$$

when

$$\delta^*(y) = \left\{i: \lambda^*(i, y) = \frac{f_{X^*Y}(i, y)}{f_{X^*}(i) f_Y(y)} > \lambda_o\right\}. \tag{7.2}$$

Moreover the solution of our new problem is exactly of the same form as that of our original problem with the exception that X is replaced by the (discrete) variable X^*. Thus where the information numbers based on the joint distribution of X and Y appeared relevant in the original problem, the somewhat reduced versions based on the joint distribution of X^* and Y are relevant in the current problem. The current problem refers to the more realistic one using piles because one cannot afford to examine each X_i individually to see if it belongs in $\delta(y)$ and deserves to be compared with y.

It is likely that a somewhat larger information number (possibly $I_K(f_{X^*Y})$) would be

relevant for the variation of this problem where the piles may be searched in descending order of $\lambda(X^*, Y)$ and one stops searching as soon as a single matching X for Y is found.

8. Metric on $\mathcal{X}$. Two somewhat different situations may exist in the selection of the piles. In one case X_i^* itself is determined by a vector of random variables, and the computation of f_{X^*Y} is relatively simple and does not require the formulae of the preceeding section. For example, X^* could represent a vector of 10 ones and zeros corresponding to the ten fingers, a one indicating a *loop* (one type of fingerprint), and a zero a nonloop. In such a case the value of y might lead to the automatic disqualification, as candidates, of any point in $\mathcal{X}$ for which there is more than one discrepancy on the type (loop or nonloop) among the ten fingers.

In this case, the decision to include $X^* \in \delta^*(y)$ requires no detailed computation and is practically cost free. Given $Y = y$, a computer could be guided to the appropriate data bans or piles directly.

A second situation is where there is no very simple description of $\mathcal{X}_i$ in terms of a low-dimensional, well-understood vector. In this case it seems plausible to group together X_j and $X_{j'}$ in the same $\mathcal{X}_i$ if they are close to each other in an appropriate sense. One measure of closeness or distance between x_1 and x_2 which suggests itself is

$$D(x_1, x_2) = -\log\left[\int \frac{f_{X_1Y}^{1/2}(x_1, y) f_{X_2Y}^{1/2}(x_2, y)}{f_{X_1}^{1/2}(x_1) f_{X_2}^{1/2}(x_2)}\, dy\right]. \tag{8.1}$$

This measure may be arrived at from several points of view. First we would like to regard x_1 and x_2 as close if $\lambda(x_1, Y)$ and $\lambda(x_2, Y)$ tend to both be large or both be small for the same values of Y. A measure of how well $\lambda(x_1, Y)$ *and* $\lambda(x_2, Y)$ are correlated or rather how well their square roots are correlated is

$$\int \lambda^{1/2}(x_1, y)\lambda^{1/2}(x_2, y) f_Y(y)\, dy$$

which is simply $\exp[-D(x_1, x_2)]$. Alternatively we may regard D as coming from the Bhattacharya distance between $f_{Y|X}(\cdot \mid x_1)$ and $f_{Y|X}(\cdot \mid x_2)$, the conditional distributions of Y given $X = x_1$ and $X = x_2$. The author has not yet had experience with any large scale implementation of such a measure. There seem to be several ways of going about it and technical problems may arise.

First, to measure D theoretically may be difficult. Presumably one may estimate D empirically by sampling from the Y distribution.

One could select the piles $\mathcal{X}_i$, which may be regarded as clusters of X's, by sampling the large population and using the points in the sample as representative centers of the $\mathcal{X}_i$. Every element of $\mathcal{X}$ could be assigned to the center to which it is closest.

Alternatively one could use the sample to estimate the distribution of $D(X_1, X_2)$ and use as centers only X_i which are far from other points sampled.

Finally one could use the distances $D(X_1, X_2)$ to do a cluster analysis based on these distances. This cluster analysis differs from the usual one in that the set of entries is very large and the number of clusters would be quite large.

It would seem advisable to obtain some experience with several approaches on a few examples before setting up an elaborate theory. It should also be pointed out that one may anticipate coming across situations such as in fingerprints where there is a substantial amount of information tightly packed into a few unspecified components of X and where there is a large amount of information loosely spread around through most of the components of X. In this latter case we may have a very large nugget of information (such as an identifying scar) which is located in an unpredictable part of X. Then it would be desirable to find a few cleverly selected features, i.e., functions of X, which summarize a great deal of effective information in compact form.

9. Two-stage search. In this section we elaborate the method described in Section 7 to a two-stage procedure. The set $\mathcal{X}$ is partitioned into disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_M$ called

piles. Each pile $\mathscr{X}_i$ is partitioned into disjoint subsets $\mathscr{X}_{ij}$, $j = 1, 2, \cdots, M_i$, $i = 1, 2, \cdots, M$. These subsets are called *bins*. There are M_{ij} elements in bin $\mathscr{X}_{ij}$. During the first stage each pile is examined and a decision is made on whether or not to search it. If it is to be searched, then a similar choice is made for each bin in that pile. Each element of a bin that is searched is matched against Y.

During Stage 1, a cost of $c_1 M$ is incurred in examining the M piles where c_1 is the cost for examining a given pile. In Stage 2, a cost of c_2 is incurred for each bin in a pile $\mathscr{X}_i$ which is searched. Finally, when a bin is searched, each element of the bin is compared with Y at a cost of c_3 per element. Let

$$\tilde{\lambda}_1 = c_1 M/k\pi, \qquad \tilde{\lambda}_{2i} = c_2 M_i/k\pi \qquad \text{and} \quad \tilde{\lambda}_3 = c_3 N/k\pi,$$

and let

$$X^* = i \text{ if } X \in \mathscr{X}_i \qquad \text{and} \quad X^{**} = (i, j) \text{ if } X \in \mathscr{X}_{ij}.$$

Let C be the expected cost of using the two-stage search. Since $k\pi$ is the expected cost of not searching the library at all, $(k\pi)^{-1}C$ represents the factor by which this basic cost is multiplied when using our scheme. In a crude way we would expect that, except for the examination costs, the use of X^* would lead to a factor R^* similar to that of our previous theory and the use of X^{**} in the second stage to another factor R^{**}. Then $(k\pi)^{-1}C$ would correspond roughly to R^*R^{**} except for the influence of $\tilde{\lambda}_1$, $\tilde{\lambda}_{2i}$, and $\tilde{\lambda}_3$. Incidentally by selecting M and M_i suitably, $\tilde{\lambda}_1$ and $\tilde{\lambda}_{2i}$ may be made small compared with 1. However $\tilde{\lambda}_3$ is ordinarily quite large.

It would be rather optimistic to expect that there would be no loss of efficiency in this two-stage process and hence it would be somewhat surprising if there were no degradation in using two stages. If there is a loss of efficiency, does it matter in what stage some specific information is presented? Suppose for example that X^{**} corresponds to two independent vectors and X^* can be chosen to be either one of these. Is there a simple rule for deciding which is preferable?

These questions are addressed in this and the next sections. Here we present a crude estimate of $(k\pi)^{-1}C$ which suggests an effect $\tilde{\lambda}_3^{1/2} R^*R^{**}$ and a modification. Reasons are given to prefer, for Stage 1, information which is more precise and more compressed. In Section 10, the results of this section are compared with some Monte Carlo computations in a few simple cases.

During Stage 1 a cost of $c_1 M$ is incurred in considering the M bins. The conditional cost given $Y = y$ is

$$C_1(y) = c_1 M + \sum_{i=1}^{M} \min[k\pi P(X \in \mathscr{X}_i \mid Y = y),\ C_2(i, y)]$$

where $C_2(i, y)$, to be considered next, is the conditional cost of searching the ith pile $\mathscr{X}_i$ given $Y = y$. Then

$$C_2(i, y) = C_2 M_i + \sum_{j=1}^{M_i} \min[k\pi P(X \in \mathscr{X}_{ij} \mid Y = y),\ c_3 N P(X \in \mathscr{X}_{ij})].$$

Ultimately we are interested in

$$\begin{aligned} C &= E[C_1(y)] \\ &= c_1 M + \int \sum_{i=1}^{M} \min[k\pi f_{X^* \mid Y}(i \mid y),\ C_2(i, y)] f_Y(y)\, dy. \end{aligned}$$

Factoring out $k\pi$ from C and C_2 yields

$$(k\pi)^{-1}C = \tilde{\lambda}_1 + \int \sum_{i=1}^{M} \min[f_{X^* \mid Y}(i \mid y),\ (k\pi)^{-1} C_2(i, y)] f_Y(y)\, dy \tag{9.1}$$

where

$$(k\pi)^{-1}C_2(i, y) = \tilde{\lambda}_{2i} + \sum_{j=1}^{M_i} \min[f_{X^{**} \mid Y}((i, j) \mid y),\ \tilde{\lambda}_3 f_{X^{**}}(i, j)]. \tag{9.2}$$

Then

$$(k\pi)^{-1}C_2(i, y) = \tilde{\lambda}_{2i} + f_{X^* \mid Y}(i \mid y) A_{21}(i, y), \tag{9.3}$$

where

$$A_{21}(i, y) = \sum_{j=1}^{M_i} \min[f_{X^{**}|X^*Y}((i, j) \mid i, y), \frac{\tilde{\lambda}_3}{\lambda_1(i, y)} f_{X^{**}|X^*}((i, j) \mid i)] \tag{9.4}$$

and

$$\lambda_1(i, y) = \frac{f_{X^*Y}(i, y)}{f_{X^*}(i) f_Y(y)} \tag{9.5}$$

is the likelihood-ratio for testing f_{X^*Y} versus $f_{X^*}f_Y$. Given $X^* = i$ and $Y = y$, A_{21} represents the minimum of $\alpha + \lambda\beta$ for testing $f_{X^{**}|X^*Y}(\cdot \mid i, y)$ versus $f_{X^{**}|X^*}(\cdot \mid i)$ with $\lambda = \tilde{\lambda}_3/\lambda_1(i, y)$. Then A_{21} can be bounded above by

$$A_{22}(i, y) = \inf_{0 \le t \le 1} [\sum_{j=1}^{M_i} f_{X^{**}|X^*Y}^{1-t} f_{X^{**}|X^*}^{t} \tilde{\lambda}_3^t \lambda_1^{-t}] \tag{9.6}$$

and, using the special case $t = \frac{1}{2}$, by

$$\begin{aligned} A_{23}(i, y) &= \tilde{\lambda}_3^{1/2} \lambda_1^{-1/2} \sum_{j=1}^{M_i} f_{X^{**}|X^*Y}^{1/2} f_{X^{**}|X^*}^{1/2}, \\ &= \tilde{\lambda}_3^{1/2} \lambda_1^{-1/2} R_B^{**}(i, y) \end{aligned} \tag{9.7}$$

where R_B^{**}, the sum in the expression on the right of (9.7), will be called the *Bhattacharya factor* for Stage 2.

In view of our comments in Section 3 these bounds, and A_{22} in particular, could be regarded as approximations bounding the reducing factor R^{**} due to the information in the second stage. A_{23} would constitute a reasonable approximation of A_{21} if the minimizing value of t in (9.6) were close to ½. The use of the Bhattacharya related bound A_{23} is particularly convenient in the following coarse estimate for $A_1(y)$ defined by

$$(k\pi)^{-1} C_1(y) = \tilde{\lambda}_1 + A_1(y). \tag{9.8}$$

Then

$$\begin{aligned} A_1(y) &\le \sum_{i=1}^{M} \min[f_{X^*|Y}, \tilde{\lambda}_{2i} + \tilde{\lambda}_3^{1/2} R_B^{**} f_{X^*|Y} \lambda_1^{-1/2}], \\ A_1(y) &\le \sum_{i=1}^{M} \min[f_{X^*|Y}, \tilde{\lambda}_{2i} + \tilde{\lambda}_3^{1/2} R_B^{**} f_{X^*|Y}^{1/2} f_{X^*}^{1/2}]. \end{aligned} \tag{9.9}$$

Ignoring the $\tilde{\lambda}_{2i}$ term in the bound and treating R_B^{**} as a constant we would have

$$\begin{aligned} A_1(y) &\le \sum_{i=1}^{M} f_{X^*|Y}^{1-t} \tilde{\lambda}_3^{t/2} (R_B^{**})^t f_{X^*|Y}^{t/2} f_{X^*}^{t/2}, \\ A_1(y) &\le \tilde{\lambda}_3^{t/2} (R_B^{**})^t \sum_{i=1}^{M} f_{X^*|Y}^{1-t/2} f_{X^*}^{t/2} \quad \text{for} \quad 0 \le t \le 1. \end{aligned}$$

In particular, for $t = 1$

$$A_1(y) \le \tilde{\lambda}_3^{1/2} R_B^{**} R_B^{*}(y) \tag{9.10}$$

where

$$R_B^{*}(y) = \sum_{i=1}^{M} f_{X^*|Y}^{1/2} f_{X^*}^{1/2} \tag{9.11}$$

is the *Bhattacharya factor* for the information in Stage 1.

Let us examine the "results" (9.9) and (9.10), what they state, the implicit assumptions in their derivations, how they can be modified and implications that they suggest.

If we think of R_B^{**} as constant, (9.9) claims the relevance, in Stage 1, of

$$B_1 = \sum_{i=1}^{M} \min[f_{X^*|Y}, \theta_{1i} + \theta_2 f_{X^*|Y}^{1/2} f_{X^*}^{1/2}] \tag{9.12}$$

where θ_{1i} corresponds to $\tilde{\lambda}_{2i}$ and θ_2 to $\tilde{\lambda}_3^{1/2} R_B^{**}$. (As an approximation rather than as a bound, one might prefer a more sensitive approximation suggested by (3.12), i.e., to replace R_B^{**} in (9.9) by a quantity of the order of $n_2^{-1/2} R_B^{**}$ where n_2 is the number of informative components of X^{**} given X^*.)

Equation (9.10) seems much cruder. It acts as though the two-stage effect is the product of the single state effects or that the combined effect is expressed in terms of the sum of the Bhattacharya information of the two stages.

These claims are subject to several major qualifications. We have treated $R_B^*(X^*, Y)$, and expect to treat $R_B(Y)$, as constants, assuming that these terms have relatively small coefficients of variations. We have ignored $\tilde{\lambda}_{2i}$ in obtaining (9.10) from (9.9). Moreover if $\tilde{\lambda}_3^{1/2} R_B^*$ is large, it is likely that the right-hand side of (9.10) can be replaced by a lower value if we use $t < 1$ in the inequality leading to (9.10). But if we use $t < 1$, then the effect of R_B^* is diminished. This may be relevant to the question raised earlier about how to distribute information in the two stages. It suggests putting more information in Stage 1 if the $\tilde{\lambda}_{2i}$ effect is large.

Since the $\tilde{\lambda}_{2i}$ effect appears in (9.9) for each pile that is examined, we wish to reduce the number of piles that should be examined for each Y. This suggests that not only should the first stage be informative, but the information should be of good quality where each bit of stored information contains small associated errors, and hence this information is compressed. In the next section we shall elaborate slightly on this point.

10. Two-stage examples. We supplement and test the crude approximations of Section 9 on two-stage search by examining a few examples. Let Z_{ij}, $1 \le i \le r$, $1 \le j \le n_i$ be independent Bernoulli random variables where

$$P(Z_{ij} = 1) = p_i \qquad P(Z_{ij} = 0) = 1 - p_i.$$

Let X_{ij} and Y_{ij} be related Bernoulli variables (observations on Z_{ij}) where

$$P(X_{ij} = 0 \mid Z_{ij} = 1) = P(Y_{ij} = 0 \mid Z_{ij} = 1) = \epsilon_1$$

and

$$P(X_{ij} = 1 \mid Z_{ij} = 0) = P(Y_{ij} = 1 \mid Z_{ij} = 0) = \epsilon_o.$$

Table 10.1 presents 3 sets of p, ϵ_1, ϵ_o, and the corresponding values of I_B, I_o and t_o abstracted from a larger version of Table 4.1. Note that the third Bernoulli random variable is much less informative than the other two because of the relatively large error probabilities ϵ_1 and ϵ_o.

These parameter sets will define the distributions of Z_{ij}, X_{ij} and Y_{ij} for $1 \le i \le 3$ in a series of examples. An *example* will consist of a two-stage problem where each stage corresponds to one of several *observation vectors*, each of which corresponds to a vector of $n = n_1 + n_2 + n_3$ observations with n_i on Z_{ij} for $1 \le i \le 3$. By taking successive examples where the observation vectors of the two stages are interchanged we may observe conditions under which certain allocations of information for the two stages are preferred to others.

We shall use a Monte Carlo simulation to evaluate the effect of two stages. To help assess the crude approximations of Section 9 we supplement the two-stage simulation in Table 10.3 by simulations of the effects of each of the stages separately in Table 10.2.

To be more specific, Equation (9.4) suggests that

$$A_{21}(i, y) \approx B_2(y, \theta) = \sum_x \min[f_{X|Y}(x \mid y), \theta f_X(x)] \tag{10.1}$$

where X and Y correspond to the observation vector of Stage 2 and θ corresponds to $\tilde{\lambda}_3/\lambda_1(i, y)$ which depends on Stage 1. Our theory tells us that $EB_2(Y, \theta) = \theta^{1/2} R_2(\theta)$ where $R_2(\theta) \le R_B = \exp(-I_B)$ is the Bhattacharya factor for the observation vector of Stage 2. (A less crude approximation would suggest replacing B_B by $R_B/n^{1/2}$ though this is not very appropriate for

TABLE 10.1
Information numbers (as in Table 4.1)

i	p	ϵ_1	ϵ_o	I_B	I_o	t_o
1	0.5	0.002	0.002	0.2872	0.3167	0.6773
2	0.3	0.002	0.005	0.2174	0.2306	0.6172
3	0.5	0.050	0.100	0.0842	0.0845	0.5304

TABLE 10.2

First stage and second stage effects in two-stage problems

Observation vector							Second stage				First stage				
n_1	n_2	n_3	n	I_B	R_B	$n^{-1/2}R_B$	θ	$\bar{B}_2$	s_{B_2}	$\theta^{-1/2}\bar{B}_2$	θ_1	θ_2	$\theta_2 R_B$	$\bar{B}_1$	s_{B_1}
6	6	0	12	3.03	0.05	0.01	1.00	0.01	0.00	0.008	0.02	0.10	0.00	0.08	0.00
							4.48	0.03	0.02	0.014		0.44	0.02	0.09	0.00
							20.09	0.06	0.02	0.014		1.30	0.06	0.11	0.01
							90.02	0.11	0.04	0.012		2.84	0.14	0.14	0.02
											0.01	0.44	0.02	0.08	0.02
											0.05	0.44	0.02	0.12	0.00
6	4	2	12	2.76	0.06	0.02	1.00	0.01	0.01	0.015	0.02	0.10	0.01	0.13	0.00
							4.48	0.04	0.01	0.017		0.44	0.03	0.14	0.00
							20.09	0.08	0.02	0.018		1.30	0.08	0.16	0.00
							90.02	0.19	0.08	0.020		2.84	0.18	0.21	0.03
											0.01	0.44	0.03	0.10	0.00
											0.05	0.44	0.03	0.23	0.00
6	2	0	8	2.16	0.11	0.04	1.00	0.03	0.01	0.033	0.02	0.10	0.01	0.06	0.00
							4.48	0.06	0.01	0.027		0.44	0.05	0.09	0.00
							20.09	0.14	0.05	0.031		1.30	0.15	0.14	0.02
							90.02	0.49	0.23	0.052		2.84	0.33	0.25	0.05
											0.01	0.44	0.05	0.08	0.00
											0.05	0.44	0.05	0.12	0.00
4	4	0	8	2.02	0.13	0.05	1.00	0.03	0.01	0.033	0.02	0.10	0.01	0.07	0.00
							4.48	0.07	0.02	0.032		0.44	0.06	0.09	0.01
							20.09	0.17	0.11	0.039		1.30	0.17	0.15	0.04
							90.02	0.53	0.34	0.056		2.84	0.38	0.27	0.09
											0.01	0.44	0.06	0.08	0.01
											0.05	0.44	0.06	0.12	0.01
2	4	6	12	1.95	0.14	0.04	1.00	0.04	0.01	0.039	0.02	0.10	0.01	0.36	0.01
							4.48	0.10	0.04	0.047		0.44	0.06	0.38	0.01
							20.09	0.23	0.08	0.051		1.30	0.19	0.41	0.02
							90.02	0.46	0.15	0.048		2.84	0.40	0.48	0.04
											0.01	0.44	0.06	0.31	0.01
											0.05	0.44	0.06	0.59	0.01
2	6	0	8	1.88	0.15	0.05	1.00	0.04	0.01	0.043	0.02	0.10	0.02	0.07	0.00
							4.48	0.09	0.04	0.042		0.44	0.07	0.10	0.01
							20.09	0.26	0.17	0.058		1.30	0.20	0.18	0.05
							90.02	0.70	0.32	0.074		2.84	0.40	0.33	0.11
											0.01	0.44	0.07	0.09	0.01
											0.05	0.44	0.07	0.13	0.01
4	2	0	6	1.58	0.21	0.08	1.00	0.05	0.01	0.046	0.02	0.10	0.02	0.06	0.00
							4.48	0.11	0.04	0.051		0.44	0.09	0.11	0.01
							20.09	0.39	0.18	0.088		1.30	0.27	0.22	0.04
							90.02	0.96	0.13	0.102		2.84	0.58	0.41	0.09
											0.01	0.44	0.09	0.10	0.01
											0.05	0.44	0.09	0.14	0.01
3	3	0	6	1.51	0.22	0.09	1.00	0.05	0.01	0.054	0.02	0.10	0.02	0.06	0.00
							4.48	0.15	0.06	0.069		0.44	0.10	0.12	0.02
							20.09	0.55	0.29	0.124		1.30	0.29	0.25	0.06
							90.02	0.92	0.19	0.098		2.84	0.63	0.48	0.13
											0.01	0.44	0.10	0.11	0.02
											0.05	0.44	0.10	0.15	0.02
2	4	0	6	1.44	0.24	0.10	1.00	0.06	0.02	0.063	0.02	0.10	0.02	0.07	0.00
							4.48	0.18	0.09	0.086		0.44	0.10	0.13	0.02
							20.09	0.63	0.31	0.141		1.30	0.31	0.27	0.07
							90.02	0.96	0.15	0.101		2.84	0.67	0.54	0.16
											0.01	0.44	0.10	0.12	0.02
											0.05	0.44	0.10	0.16	0.02

our observation vectors where observations are discrete and not identically distributed, and n is relatively small.) As part of Table 10.2 we list, for each of nine observation vectors, the average $\bar{B}_2$ and the standard deviation s_{B_2} of 25 observations on $B_2(Y, \theta)$ as well as R_B, $R_B/n^{1/2}$ and $\theta^{-1/2}\bar{B}_2$.

We see from the relative stability of $\theta^{-1/2}\bar{B}_2$ close to $R_B/n^{1/2}$ that

$$E[B_2(Y, \theta)] \approx \theta^{1/2}R_B/n^{1/2} \tag{10.2}$$

serves as a good approximation over the range of vectors studied and we may reasonably regard $R_2(\theta)$ as a constant R_2 which may be estimated by $\overline{B_2(Y, 1)}$ or by $R_B/n^{1/2}$.

The approximation above was somewhat better than expected even though it degenerates a bit for large θ. I would expect that for high dimensional observation vectors, a multiplicative factor of order $O(1)$ might be useful to maintain a good approximation.

Equation (9.12) suggests the relevance of

$$B_1(y, \theta_1, \theta_2) = \sum_x \min[f_{X|Y}(x \mid y), \theta_1 + \theta_2 f_{X|Y}^{1/2}(x \mid y) f_X^{1/2}(x)] \tag{10.3}$$

where X and Y correspond to the observation vector of the first stage, θ_1 corresponds to $\tilde{\lambda}_2$, and θ_2 to $\tilde{\lambda}_3^{1/2}R_2$. Then one should anticipate that

$$A_1(y) \approx B_{11} = E\{B_1(Y, \tilde{\lambda}_2, \tilde{\lambda}_3^{1/2}E[B_2(Y, 1)])\} \tag{10.4}$$

or

$$A_1(y) \approx B_{12} = E\{B_1(Y, \tilde{\lambda}_2, \tilde{\lambda}_3^{1/2}R_{B2}n_2^{-1/2})\} \tag{10.5}$$

where R_{Bi} and n_i are the Bhattacharya factor and number of components for the observation vector of Stage i. Finally Equation (9.10) suggests

$$A_1(y) \approx B_{13} = \tilde{\lambda}_3^{1/2}R_{B1}R_{B2}. \tag{10.6}$$

These anticipations are tested in Table 10.3 but EB_1 is estimated in the remainder of Table 10.2 which lists the average $\bar{B}_1$ and standard deviation s_{B_1} of 25 observations on $B_1(Y, \theta_1, \theta_2)$ and compares $\bar{B}_1$ with $\theta_2 R_B$ which is suggested by the "derivation" of (9.10).

In the Stage 1 part of Table 10.2 we see that observation vectors with comparable I_B differ considerably in effect on $\bar{B}_1$. First of all, for comparable I_B, $\bar{B}_1$ tends to be larger when $n_3 > 0$ than when $n_3 = 0$. When $n_3 > 0$, $\bar{B}_1$ is relatively flat in θ_2, and grows very rapidly with θ_1 while $\bar{B}_1$ is additive in θ_1 if $n_3 = 0$.

The observation vector (6, 6, 0) does not do as well as (3, 3, 0), which carries half as much information, if θ_2 is small.

The estimate $\theta_2 R_B$ is rather poor since $\theta_2^{-1}\bar{B}_1$ is not very stable.

The poor quality of the observation vectors for which $n_3 > 0$ may be attributed to the fact that more piles need searching when these vectors are used in Stage 1. Note that if (6, 6, 0) is used only one pile is searched (that where $X^* = Y^*$), but it is more likely that an error, $X^* \neq Y^*$, will occur than if (3, 3, 0) is used. This explains why (3, 3, 0), though less informative, is better for small θ_2. This apparently paradoxical result is based on our restriction that each pile be examined separately to see if it should be searched and with (6, 6, 0) there may be more bins and piles to examine. (Of course we are implicitly assuming that behavior of B_1 does reflect on the overall two-stage behavior.)

In Table 10.3 we present the average $\bar{A}_1$ and standard deviation s_{A_1} of 25 observations on $A_1(Y)$ for several two-stage examples. The second example is derived from the first by interchanging the observation vectors of the two stages. The fourth is similarly obtained from the third. In each example $\bar{A}_1$ is compared with three quantities. These are estimates from Table 10.2 of B_{11}, B_{12} and B_{13}.

In the first two examples we see clearly that the order of stage is important. Consequently, B_{13} which fails to take this order into consideration is a relatively useless approximation to $\bar{A}_1$.

Comparing the fifth example with the first and the last two examples with each other, we see again that additional information in Stage 1 is potentially harmful. For the classical

TABLE 10.3
Two-stage simulation

$\bar{A}_1$ *and* s_{A_1} *based on* 25 *observations*

$\tilde{\lambda}_2 = 0.02 \quad \tilde{\lambda}_3 = 1{,}000$

Observation vectors														
Stage 1				Stage 2										
n_1	n_2	n_3	R_{B1}	n_1	n_2	n_3	R_{B2}	$\bar{A}_1$	s_{A_1}	θ_{21}	θ_{22}	B_{11}	B_{12}	B_{13}
2	4	6	0.14	6	2	0	0.11	0.59	0.01	1.03	1.29	0.40	0.41	0.52
6	2	0	0.11	2	4	6	0.14	0.13	0.03	1.22	1.30	0.14	0.14	0.52
3	3	0	0.22	6	6	0	0.05	0.11	0.02	0.27	0.44	0.08	0.12	0.33
6	6	0	0.05	3	3	0	0.22	0.13	0.02	1.72	2.84	0.13	0.14	0.33
2	4	0	0.24	6	2	0	0.11	0.25	0.13	1.03	1.29	0.16	0.27	0.86
4	4	0	0.13	6	6	0	0.05	0.10	0.01	0.27	0.44	0.08	0.09	0.20
6	4	2	0.06	6	6	0	0.05	0.23	0.03	0.27	0.44	0.13	0.14	0.10

θ_{21}, B_{11} and B_{12} are based in part on Table 10.2.

$B_{11} = \bar{B}_1(Y, \tilde{\lambda}_2, \theta_{21}) \qquad \theta_{21} = \tilde{\lambda}_3^{1/2}\overline{B_2(Y, 1)}$

$B_{12} = \bar{B}_1(Y, \tilde{\lambda}_2, \theta_{22}) \qquad \theta_{22} = \tilde{\lambda}_3^{1/2} R_{B2} n_2^{-1/2}$

$B_{13} = \tilde{\lambda}_3^{1/2} R_{B1} R_{B2}$

statistician it must be emphasized that this effect exists because the restrictions of our search behavior do not permit us to do the equivalent of ignoring data in statistical problems.

In conclusion we have the following remarks.

1. Generally it is most desirable to have more information in Stage 1 when $\tilde{\lambda}_3$ is large.

2. If $\tilde{\lambda}_2$ is large then the quality of the information in Stage 1 is important. Poor quality implies the need to search more piles and examine more bins. Then a large $\tilde{\lambda}_2$ is costly. Generally, well compressed data are of good quality. However, as the observation vectors (6, 6, 0) and (3, 3, 0) indicate, for small $\tilde{\lambda}_3$, the increased likelihood of a mismatch in Stage 1 if we use more information makes that information less desirable even though it may be thought of as equally compressed as (3, 3, 0). One might argue that the $n^{-1/2}$ factor implies that the information in (6, 6, 0) is actually less compressed but I doubt that this point is important. The large atoms of the limited discrete data in our examples may play a more important role.

3. Tables 10.2 and 10.3 support the relevance of EB_2 and $n_2^{-1/2}R_{B2}$ for stage 2, $E(B_1)$ for Stage 1, and $B_1(y, \tilde{\lambda}_2, \tilde{\lambda}_3^{1/2} n_2^{-1/2} R_{B2})$ for the two-stage examples. The usefulness of (9.10) seems to be extremely limited.

4. Computations based on Tables 10.2 and 10.3 support the suggestion that it is when $\tilde{\lambda}_2$ is large that it is most valuable to put more and better compressed data in Stage 1.

Acknowledgments. I wish to thank Peter Elias for the benefit of several discussions and comments and for observing that in Table 5.1, I_o more than doubles as k goes from 2 to 4.

REFERENCES

[1] BHATTACHARYA, A. (1943). On a measure of divergence between two statistical populations. *Bull Calcutta Math. Soc.* **35** 99–109.

[2] BLACKWELL, D. and HODGES, J. L. (1959). The probability in the extreme tail of a convolution. *Ann. Math. Statist.* **30** 1113–1120.

[3] CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23** 493–507.

[4] CHERNOFF, H. (1972). *Sequential Analysis and Optimal Design.* SIAM, Philadelphia, Pa.

[5] CHERNOFF, H. (1977). Some applications of a method of identifying an element of a lar; e multidi-

mensional population. In *Multivariate Analysis—IV.* (ed. P. R. Krishnaiah) 445–456. North Holland Pub. Co.

[6] CRAMÉR, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités Sci. Indust.* **F36** Paris.

[7] MATUSITA, K. (1964). Distance and decision rules. *Ann. Inst. Statist. Math.* **16** 305–316.

[8] SUNTER, A. B. and FELLEGI, I. P. (1967). An optimal theory of record linkage. In *Proc. 36th Session Internat. Statist. Inst.* 809–835. Sydney.

DEPARTMENT OF MATHEMATICS, ROOM 2-381
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139