International Book Company: New York, (1977).
[4] Pierce, John R., An Introduction to Information Theory: Symbols, Signals, and Noise, Second Edition, Dover Publications, Inc.: New York, (1980).
[5] Shannon, C.E., Bell System Technical Journal, 27, 379 and 623 (1948).
[6] Malissa, Conveners H., and J. Rendl, Z. Anal. Chem., 272, 1 (1974) English version: I.L. Marr, Talanta, 22, 597 (1975).
[7] Eckschlager, Karel, and Vladimir Stepanek, "Information Theory as Applied to Chemical Analysis", John Wiley & Sons: New York, (1979).
[8] Eckschlager, Karel, and Vladimir Stepanek, Anal. Chem., 54(11), 1115 (1982).
[9] Ritter, S.L.; S.R. Lowery, H.B. Woodruff and T.L. Isenhour, Anal. Chem., 48(7), 1027 (1976).
[10] Grys, Stanislaw, Z. Anal. Chem., 273, 177 (1975).
[11] Eckschlager, Karel, Anal. Chem., 49(8), 1265 (1977).
[12] Danzer, K., Z. Chem., 15, 326 (1975).
[13] Danzer, Klaus, and Karel Eckschlager, Talanta, 25, 725 (1978).
[14] Kaiser, H., Anal. Chem., 42(2), 24A (1970).
[15] Fitzgerald, J.J., and J.D. Winefordner, Rev. Anal. Chem., 2(4), 229 (1975).
[16] Yost, R.A., Spectra, 9(4), 3 (1983).
[17] Fetterolf, D.D., and R.A. Yost, Int. J. Mass Spectrom. Ion Processes, 62, 33 (1984).
[18] Boudreau, P.A., and S.P. Perone, Anal. Chem., 51(7), 811 (1979).
[19] Cleij, P., and A. Dijkstra, Fresenius Z. Anal. Chem, 298, 97 (1979).
[20] Brillouin, L., "Science and Information Theory", Academic Press: New York, (1962).
[21] Cleij, P., and A. Dijkstra, Fresenius Z. Anal. Chem, 294, 361 (1979).
[22] Dupuis, Foppe, and Auke Dijkstra, Anal. Chem., 47(3), 379 (1975).
[23] Eskes Arie, Foppe Dupuis, Auke Dijkstra, Henri De Clercq, and Desire L. Massart, Anal. Chem., 47(13), 2168 (1975).
[24] van Marlen, Geert, and Auke Dijkstra, Anal. Chem., 48(3), 595 (1976).
[25] Eckschlager, K., Coll. Czech. Chem. Commun., 41, 2527 (1976).
[26] Perone, S.P., ACS Symposium Series, 265 (Comput. Lab.), 99 (1984).
[27] Burgard, D., and S.P. Perone, Anal. Chem., 50(9), 1366 (1978).
[28] Byers, W. Arthur, and S.P. Perone, Anal. Chem., 55(4), 615 (1983).
[29] Byers, W. Arthur, B.S. Freiser, and S.P. Perone, Anal. Chem., 55(4), 620 (1983).
[30] Barnes, Freddie, and S.P. Perone, unpublished results, personal communication from Freddie Barnes, September 1984.
[31] Bard, Allen J., and Larry R. Faulkner, "Electrochemical Methods, Fundamentals and Applications", John Wiley & Sons: New York, (1980).

DISCUSSION

of the Perone-Ham paper, Measurement and Control of Information Content in Electrochemical Experiments

Herman Chernoff
Statistics Center
Massachusetts Institute of Technology

The Shannon theory of information has had a profound impact in science and technology. Shannon defined information in terms of the reduction of uncertainty which, in turn, was measured by entropy. He was concerned mainly with the use of information to measure the ability to transmit data through noisy channels, i.e., channel capacity.

Statisticians have developed other, somewhat related, notions of information. In statistical theory, the major emphasis has been on how well experimental data help to achieve the goals in the classical statistical problems of estimation and hypothesis testing. These measures serve two useful functions. They serve to set a standard for methods of data analysis, methods whose efficiencies are measured in terms of the proportion of the available information that is effectively used. They also serve to design efficient experiments.

For the problem of estimation, Fisher introduced the Fisher Information which we now define. Suppose that it is desired to estimate a parameter θ using the result of an experiment which yields the data X with the density $f(x|\boldsymbol{\theta})$. The *Fisher Information* for $\boldsymbol{\theta}$ corresponding to X is given by the matrix

$$J = I_X(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\mathbf{Y}\,\mathbf{Y}^T) \tag{1}$$

where $\mathbf{Y}$ is the *score function* defined by

$$\mathbf{Y} = \mathbf{Y}(X, \boldsymbol{\theta}) = \left.\frac{\partial[\log f_X(x|\theta)]}{\partial\theta}\right|_{x=X} \tag{2}$$

If $\boldsymbol{\theta}$ is a multidimensional vector, J is a nonnegative definite

symmetric matrix with the additive property

$$I_{X,Z}(\boldsymbol{\theta})=I_X(\boldsymbol{\theta})+I_Z(\boldsymbol{\theta}) \tag{3}$$

if X and Z are independent. As a consequence

$$I_{X,X,\ldots,X}(\boldsymbol{\theta})=nI_X(\boldsymbol{\theta})=nJ \tag{4}$$

if the left subscript refers to the independent replicaton of X, n times. For such an experiment, it has been shown, under mild regularity conditions, that the Maximum Likelihood Estimate (MLE) $\hat{\theta}$ will be approximately normally distributed with mean θ and covariance matrix J^{-1}/n for large n. Moreover, the Cramér-Rao theorem states that one cannot expect to find a reasonable estimate that does better.

Some implications of the above paragraph are illustrated by three simple examples below.

Example 1. *Mean of a Normal Distribution.*

Let X be normally distributed with mean θ and known variance σ^2, and let $X_1,X_2,\ldots,X_n$ be a sample of n independent observations on X. It is easy to show that $J_X=\sigma^{-2}$ and that the MLE $T_1=\bar{X}=n^{-1}(X_1+\ldots+X_n)$ is normally distributed with mean θ and variance σ^2/n. However, the statistician, who fears for outliers and may wish to use a more robust estimator than the sample mean, may prefer to use T_2, the sample median. It can be shown that T_2 is approximately normally distributed with mean θ and variance $\pi\sigma^2/2n$ for large n. The equation

$$\frac{\sigma^2}{n_1}=\frac{\sigma^2}{n_2}\frac{\pi}{2} \tag{5}$$

implies $n_1/n_2=2/\pi=.64$ which is a natural measure of the efficiency of T_2, indicating that, with T_1, we need only 64% of the data to achieve the same accuracy as with T_2. If the effective waste of 36% of the data seems excessive, the statistician can improve on efficiency with little sacrifice of robustness, e.g., by using the upper and lower quartiles as well as the median, or by using trimmed means.

Example 2. *Experiments With Information Matrices.*

Let $\theta=(\theta_1,\theta_2)^T$, and let X and Z be two experiments with information matrices

$$J_X=\begin{Vmatrix} 4 & 3 \\ 3 & 4 \end{Vmatrix} \text{ and } J_Z=\begin{Vmatrix} 4 & -3 \\ -3 & 4 \end{Vmatrix}.$$

It is desired to estimate θ_1 using replications of either (X,X) or (Z,Z) or (X,Z). Let J^{11} be the upper left member of J^{-1} which measures the (asymptotic) variance of $\hat{\theta}_1$, the MLE of θ_1. Then

$$J_{XX}=\begin{Vmatrix} 8 & 6 \\ 6 & 8 \end{Vmatrix}, \qquad J^{11}_{XX}=0.286,$$

$$J_{ZZ}=\begin{Vmatrix} 8 & -6 \\ -6 & 8 \end{Vmatrix}, \qquad J^{11}_{ZZ}=0.286,$$

and

$$J_{XZ}=\begin{Vmatrix} 8 & 0 \\ 0 & 8 \end{Vmatrix}, \qquad J^{11}_{XZ}=0.125.$$

This clearly indicates, that in the presence of *nuisance parameters* such as θ_2, one may squeeze much more useful information out of a combination of two equally informative experiments than by repeating one of these two or, in this case, even four times.

Example 3. *Estimate Safe Dose Level in Probit Model.*

For an experiment at does level d, the probit model attributes the probability of a response to be

$$p(d,\boldsymbol{\theta})=\Phi[(d-\mu)/\sigma] \tag{6}$$

where $\boldsymbol{\theta}=(\mu,\sigma)^T$, Φ is the standard normal cumulative distribution and the "safe" dose level to be estimated is defined as $\mu-2.87\sigma$. If one is permitted to select a sequence of n dose level, $d_1,d_2,\ldots,d_n$ with which to challenge n subjects, the optimal choice or *design*, for estimating $\mu-2.87\sigma$ can be shown to assign about 23% of the doses at level $d=\mu+1.57\sigma$ and the remaining 77% of the doses at level $d=\mu-1.57\sigma$.

This optimal design illustrates several points.

1. This design is *locally optimal*, i.e., it requires a knowledge of $\boldsymbol{\theta}$ to provide the best estimate of a function of θ. Superficially, it seems silly, for if we knew θ, we would not need to estimate it. In fact, it indicates that as data cumulates, one knows more about $\boldsymbol{\theta}$ and can sequentially use that information to provide improved experiments.

2. In this experiment, the repeated use of one dose level d_o would provide only an estimate of the function $p(d_o,\boldsymbol{\theta})$ and would yield no other useful information about $\boldsymbol{\theta}$ or $\mu-2.87\sigma$. At least two dose levels are required. What is somewhat surprising is that no more than two dose levels are required for an optimal design. A more general theorem states that if it is desired to estimate r functions of k parameters upon which the distribution of the data depend, then an optimal design can be constructed using at most $k+(k-1)+\ldots+(k-r+1)$ of the available (elementary) experiments.

3. The optimal design is not necessarily a practical one. Most investigators would be interested in using a variety of dose levels as a means of checking the basic model. Theory permits us to measure the loss of information inherent in the use of practical, but suboptimal designs, so that one can

decide on whether the loss is so extravagant that other alternatives should be considered.

We mention briefly that for testing hypotheses, there are several measures of information which are of potential use, depending on the type of problem. Perhaps the most useful measure is the *Kullback-Leibler Information* (KL)

$$I_X^*(\theta,\phi)=\int f(x|\theta)\log\left[\frac{f(x|\theta)}{f(x|\phi)}\right]dx \tag{7}$$

which measures two important aspects of the ability to use a sample of n observations on X with distribution $f(x)$, to discriminate between the hypotheses $H_1{:}f(x)=f(x|\theta)$ and $H_2{:}f(x)=f(x|\phi)$.

The Kullback-Leibler Information is additive as is the Fisher Information but it is not symmetric since, $I_X^*(\theta,\phi)$ is not generally equal to $I_X^*(\phi,\theta)$. For large samples, it is possible to find tests which, for fixed type 1 error probability $\alpha=P(\text{reject } H_1|H_1)$, have the type 2 error probability $\beta=P(\text{accept } H_1|H_2)$ approach 0 at a rate determined by I^*. We have, roughly

$$\beta^*\sim e^{-nI_X^*(\theta,\phi)} . \tag{8}$$

Another property of KL is that for optimal sequential testing as the cost c, per observation, approaches zero, the expected costs $R(\theta)$ and $R(\phi)$ associated with the sequential procedure when H_1 and H_2 are true, satisfies

$$R(\theta)\approx -c\log c/I_X^*(\theta,\phi)$$

$$R(\phi)\approx -c\log c/I_X^*(\phi,\theta) . \tag{9}$$

This implies that if we suspect H_1 is true, we should select the experiment which maximizes $I^*(\theta,\phi)$ and if we suspect that H_2 is true, we should maximize $I^*(\phi,\theta)$. Here again, as in the estimation problem, we are in a position to improve the experimental design as information cumulates, and our belief in H_1 or H_2 increases.

To return to chemical experimentation, one should point out that an experimental set up which yields vast amounts of bits of information is not very useful if the analysis of the data does not make efficient use of the data. To discriminate between two alternatives requires only one bit of *effective* information in the Shannon sense. The choice between experiments which yield 1,000 and 10,000 bits must involve how much effective information is readily available from the analysis.

Some bibliography on the uses of information in statistics is contained in Chernoff (1972).

[1] Chernoff, H., *Sequential Analysis and Optimal Design* SIAM monograph 8, SIAM, Philadelphia (1972).