CONTROL SYSTEMS LABORATORY

NOTES ON THE ESTIMATION OF INFORMATION MEASURES

Report Number R-56

May 1954

Contract DA-36-039-SC-56695

Project 8-103A, D/A Project 3-99-10-101

UNIVERSITY OF ILLINOIS · URBANA · ILLINOIS

"The research reported in this document was made possible by support extended to the University of Illinois, Control Systems Laboratory, jointly by the Department of the Army (Signal Corps and Ordnance Corps), Department of the Navy (Office of Naval Research), and the Department of the Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA-36-039-SC-56695, Project 8-103A, D/A Project 3-99-10-101."

Report Number R-56

NOTES ON THE ESTIMATION OF INFORMATION MEASURES

May 1954

Prepared by:

Albert A. Blank

Henry Quastler

CONTROL SYSTEMS LABORATORY
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS
Contract DA-36-039-SC-56695

Numbered Pages: 36

Foreword

This collection of notes is a by-product of psychological investigations. The purpose of these studies was to assess human performance in information processing. In particular, it has been attempted to measure human capabilities of acquiring and transmitting information. Measurements were obtained in terms of the Shannon-Wiener Measure of Information, and the related Measure of Transmission Rate. This implies categorization of all stimuli and responses, and estimation of the probabilities of occurrences for all possible associations of stimuli and responses. In many engineering applications, the number of categories is low; in psychological experiments, it tends to be high. For instance, in a letter-recognition experiment, there are 26 possible inputs and outputs, and 676 stimulus-response pairs. Furthermore, the probability of a given answer to a given stimulus depends also on preceding and simultaneous (neighboring) other stimuli and responses; thus, the number of distinguishable categories becomes very large. In order to associate a probability measure with every single category, a large sample is needed; the greater the precision required, the larger the sample. This can lead to an inordinate amount of labor in data taking and computing. Moreover, it seems that one cannot reach arbitrarily high precision by extending the observation to great length; it is likely that during a long series of trials, and partly as a consequence of such trials, the underlying probabilities remain not constant. Thus, it is more than a convenience to

replace the exact computation with approximating shortcuts, based on samples of moderate size.

In this laboratory, in dealing with specific aspects of these problems which arose from experimental studies, we have tried to obtain solutions of slightly greater generality than needed in the particular instance. The result are a number of techniques which have worked in some cases, and may be expected to be useful in others; they are presented in this report.

Henry Quastler

NOTES FOR THE ESTIMATION OF INFORMATION MEASURES

Table of Contents

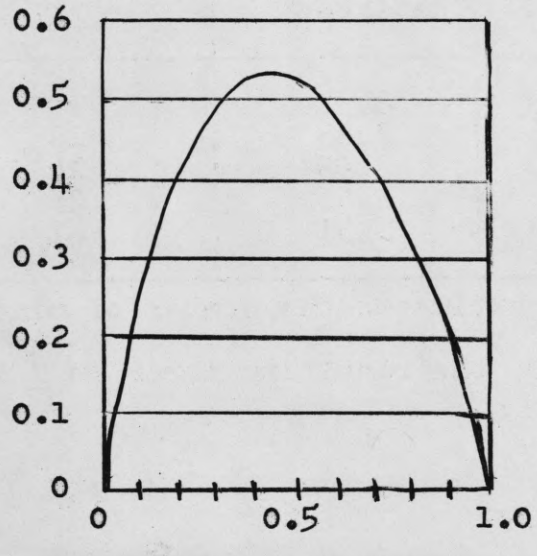
INFORMATION MEASURES IN INCOMPLETELY KNOWN SITUATIONS

Henry Quastler

The estimation of information (or uncertainty, specificity, entropy), $\underline{H}$, involves the following three operations: (1) the classification of all relevant occurrences, (2) the estimation of the probability associated with each class, and (3) the computation of an information function based on these estimates. In the ideal case, if it is known that there are exactly $\underline{r}$ classes, and that the probabilities are $\underline{p(i)}$ (where $i = 1, 2 \ldots r$; $\sum_i p(i) = 1$); then, the uncertainty, $\underline{H}$, is defined by the well-known equation

$$H = - \sum_i p(i) \log p(i).$$

In most actual situations, it is impossible to know all the alternatives or to assess accurately the probabilities associated with every single one. This could seriously impair the practical value of information measures; it is the purpose of this note to show that it does not. It will be demonstrated that a rough estimate of $\underline{H}$ is possible as soon as the major alternatives and their approximate probabilities are known.



(Fig. 1)
Function
$p \log \frac{1}{p}$

The measure of specificity is a sum of terms "p log $\frac{1}{p}$". This function rises steeply up to p = .10, which accounts for the information measure being not very sensitive to rare alternatives; and, it has a flat top for values of p between 0.20 and 0.60, which accounts for the small effect of moderate fluctuations in probability. (Fig. 1)

a - Small Effect of Rare Occurrences

The following examples will illustrate that the measure of specificity is not sensitive to rare alternatives. To begin with a hypothetical case: Suppose nine alternatives are known to account for 90% of all occurrences; if they are equiprobable, then their contribution to the measure of specificity is $9 \times \frac{1}{10} \log_2 10 = 2.99$. We now fill the remaining 10% with a varying number of equiprobable alternatives; the results are tabulated below:

TABLE I:
Effect on Uncertainty of Minor Alternatives (adding up to 10% of all occurrences).

No. of Minor Alternatives responsible for residual 10%	Probability of each	Aggregate Contribution to H	Total H if 90% of occurrences made up by 9 equiprobable alternatives	No. of equiprobable alternatives giving same total H
(Fig. 2a) 1	1/10	.33	3.32	10
(" 2b)10	1/100	.66	3.65	13
100	1/1000	1.00	3.99	16
10,000	1/100,000	1.66	4.65	25
1.000,000	1/10,000,000	2.33	5.31	41

In this situation, if one underestimates the number of minor alternatives by a factor of 100, the resulting error in H is

only about one-sixth. If only 5% of all occurrences are to be filled by unknown minor alternatives, their contribution is even less conspicuous:

TABLE II:
Contributions to Total Uncertainty by Minor Alternatives
(adding up to 5% of all occurrences)

No. of minor alternatives	Probability of each	Contribution to H
1	5/100	.22
10	5/1000	.38
1000	5/10,000	.55
10,000	5/1,000,000	.88
1,000,000	5/1,000,000,000	1.21

The difference between the extreme values in this table can be made intuitively clear by expressing uncertainties in terms of equivalent number of equiprobable alternatives (as shown in the table I). Suppose that this number is known for a single alternative accounting for 5% of the occurrences, the other 95% being distributed in any arbitrary fashion; then, if the single alternative is replaced by one million equiprobable ones, without otherwise changing the distribution, the equivalent number of equiprobable alternatives is just doubled.

An investigation by A. A. Blank furnished an impressive real example. He calculated the specificity of single English words; for particular reasons, the sample was restricted to 4-letter words. The uncertainty was obtained as

$$\hat{H} = \sum_i \frac{\nu_i}{N} \log_2 \frac{N}{\nu_i} \qquad (N = \sum_i \nu_i)$$

where ν_i is the observed frequency of the i'th 4-letter word in the Thorndyke list. He also determined the values of H obtained by successive elimination of the less frequent words. The results are shown in table III:

TABLE III:
Measure of Uncertainty for 4-letter words (data of A. A. Blank)

	no. of words	% of all words	H	%H
All 4-letter words in Thorndyke's list	1550	100.0	8.13	100
Only words with frequency $\geq$ 150	865	55.8	7.98	98.1
Only words with frequency $\geq$ 750	395	25.5	7.47	91.8
Only words with frequency $\geq$ 1550	214	13.8	6.89	84.8
Only words with frequency $\geq$ 3150	119	7.7	6.34	77.8

Thus, taking into consideration only 1/10 of all categories (which probably account for more than 1/2 of all occurrences) yields already about 4/5 of the final measure of specificity.

The examples given show that the information function is not sensitive to rare occurrences - which means that it should not be used whenever infrequent occurrences must be heavily weighed; on the other hand, it can be used successfully in situations which are not completely known.

b - Small Effect of Small Variations in Probability

Any operation which tends to average probabilities increases the uncertainty; therefore, if r alternatives have an aggregate probability P then their contribution to the

measure of uncertainty is greatest if they have all equal probabilities (P/r). We will now demonstrate that moderate deviations from equiprobability do not markedly affect the uncertainty.

Consider the simplest case of $\underline{P} = 1$ and $\underline{r} = 2$. If the two probabilities are equal, then $\underline{H} = 1$; if their ratio is 1:2, $\underline{H} = 0.92$; if the ratio is 1:3, a very considerable deviation from equality, $\underline{H}$ is still 0.81.

For larger values of $\underline{r}$, the insensitivity of $\underline{H}$ against probability distortion is still more pronounced. For instance, we may replace the 9 equiprobable major alternatives in our first example (table I) by sets of 9 alternatives with probabilities staggered arithmetically (fig. 3a) or geometrically (fig. 3b), stipulating only that the span between the extreme values should be within one order of magnitude. The resulting changes in $\underline{H}$ are quite small.

We come to the following conclusion: if a situation is analyzed to the degree that we feel we can classify 90-95% of all occurrences; and if the probabilities associated with each class are approximately known; then we are entitled to make a rough estimate of the information measure.

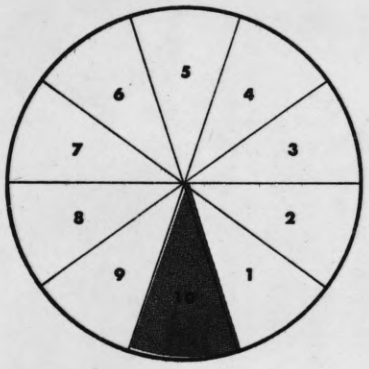


Fig. 2A 10 EQUIPROBABLE ALTERNATIVES
H = 3.33

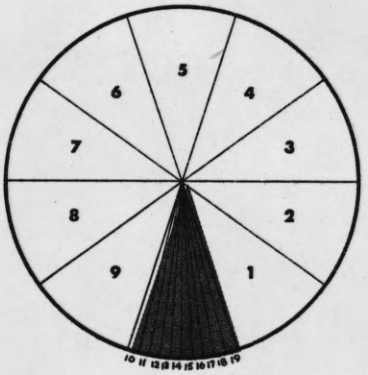


Fig. 2B 9 MAJOR, 10 MINOR ALTERNATIVES
Equivalent number of equiprobable alternatives: 13
H = 3.65

Fig. 2 Small effect of adding minor alternatives

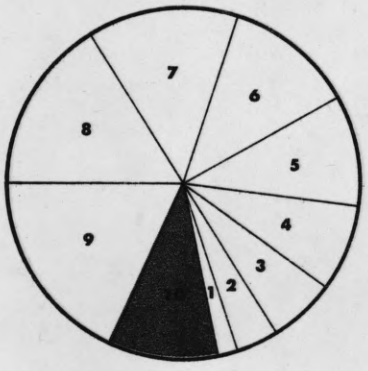


Fig. 3A EQUIVALENT NUMBER OF EQUIPROBABLE ALTERNATIVES ≈ 9
H = 3.11

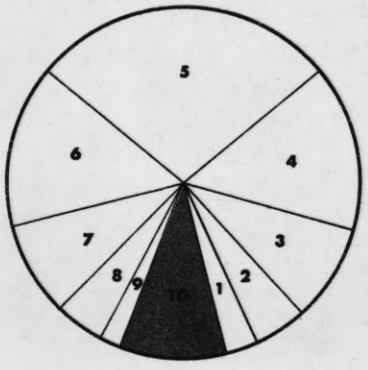


Fig. 3B EQUIVALENT NUMBER OF EQUIPROBABLE ALTERNATIVES = 8
H = 3.01

Fig. 3 Small effect of varying probabilities

EQUIVOCATION AS THE SUM OF ERROR-LOCATING AND ERROR-CORRECTING INFORMATION

H. Quastler

The estimation of information transmission is based on the estimation of probabilities for all possible input-output pairs. There are many cases, in particular those where human performances are to be assessed, where there is a host of input-output categories, more than one could ever hope to fill adequately by means of the usual experimental sampling procedures, and more than would tax the resources of a small battery of computers. For that reason, it is useful to have bounds for information transmission and equivocation. In case of doubt, the bound should be on the conservative side (which means a lower bound for transmission, an upper bound for equivocation). Ideally, an efficient bound is sought. This means that if one does make suitable allowance for all possible contingencies in the input-output table, the bound should be equal to the value sought. Such a bound for the equivocation can be constructed as follows: we imagine an auxiliary source (an "ideal observer") which, knowing both input and output, furnishes such information as is needed to reconstruct the input from the output. We will show that if most efficient coding is used, the amount of information produced by the auxiliary source becomes equal to the equivocation.

Let H(in), H(out), and H(aux) denote the uncertainties (per unit act) of input, output, and auxiliary source, respectively. Subscripted symbols denote conditional uncer-

tainties. Then (Shannon, p. 37),

$$H_{out,aux}(in) \geqslant H_{out}(in) - H(aux)$$

If it is possible to reconstruct the input completely from the output and the auxiliary message, then $H_{out,aux}(in)$ vanishes and

$$H(aux) \geqslant H_{out}(in)$$

which gives the upper bound desired. It has been shown (Shannon's theorem 10, p. 37) that it is possible to approximate $H_{out}(in)$ by H(aux) as closely as it is desired by means of efficient coding. More precisely, it is possible to correct all but an arbitrarily small amount ε of the errors in the output with a channel of capacity $H_{out}(in)$. For such an efficient code we have necessarily

$$H_{out}(in) \geqslant H(aux) \geqslant H_{out}(in) - \varepsilon$$

It follows that the uncertainty of a fully-correcting auxiliary source is an efficient bound as desired.

Shannon's theorem states what could be ideally obtained with perfect coding; it does not say how such a code is to be constructed. In the situation we are considering, the auxiliary source would have to be designed with perfect knowledge of the properties of the input and of the channel. But, it is precisely such knowledge which we try to establish by experimental tests; hence, we cannot establish an optimum behavoir for the ideal observer. What we can actually do is the

following: we begin by designing an auxiliary source which will make the equivocation vanish. The information furnished by this source is necessarily an upper bound of the amount of equivocation. Subsequently, we use any knowledge we may have about regularities in the occurrence of errors to reduce the amount of information required from the auxiliary source. This can be done in successive steps. In addition, we can try to name lower bounds for the amount of information required from the auxiliary source; this, if successful, will bracket the true value of the equivocation.

It is sometimes convenient to partition the auxiliary source into two sources, one which furnishes data needed to locate errors, and one which serves to correct errors after they are located. Accordingly, the amount of equivocation is partitioned into error-locating information of amount H(loc), and error-correcting information of amount H(cor):

$$H_{out}(in) = H(loc) + H(cor)$$

The amount of information needed for error location depends on the pattern of making errors. If this is a lawful and known pattern, then no error-locating information is needed. Suppose, for instance, that S makes a mistake exactly every 4th time; then to know the location of all errors one has to know only the location of any one false transmission (a negligible amount of information compared to the total information in a long message). Or, if it is known that S is always

wrong when his outputs are, say, of the kind "M," "P," or "Z," then no extra information will be needed to disregard these parts of his report. If, on the other hand, the commission of an error is a random event, then some information is needed to locate it. Suppose, now, that the total ensemble of errors committed includes errors which are quite lawful and predictable (provided the laws are known) and others which are random events. As only the latter need any error-locating information, it follows that the larger the fraction of random errors, the greater the amount needed for error location.

If, on the basis of prior knowledge of the S's behavior, one can extract from the output itself any indications concerning error location, then these indications can and should be used in constructing the error locating code. If it is known that all outputs are equally liable to be erroneous, then the amount of the error locating information is maximum. Therefore, if one does not know how errors are distributed in the output, he may assume them to be equiprobable and thus obtain an upper bound of the amount of information needed to supplement the S's report in order to locate the errors it contains. Deviations from equal distribution of errors have to be quite pronounced before they reduce the value of the supplementary information much below the maximum.

The following example is taken from experiments performed in this laboratory (Quastler and Wulff). Ss had the task of copying sequences of random letters on the typewriter. In a particular sample of 10,260 letters, we found 3.4% errors.

In first approximation, we assume that errors occur completely randomly; with this model, we get:

$$H(\text{loc}) = -0.034 \log_2 0.034 - 0.966 \log_2 0.966 = 0.214 \text{ bits/letters}$$

However, it was quite obvious that clusters of errors occurred more frequently than could be expected by chance grouping. Thus, in one particular sample of 10,260 letters we found:

single errors	158	times, or	1.54%
pairs	42	" "	0.41%
triplets	22	" "	0.21%
quadruplets	9	" "	0.09%
quintuplets	1	" "	0.01%

Thus, while the over-all probability of any key being wrong was 3.4%, the probability of a wrong key following a wrong key was 33.5%, or almost ten times as large. Accordingly, we introduce a second approximation, and treat the error-generating mechanism as a Markov process, with the probability of an error occurrence depending on (and only on) the success or failure of the preceding act. The Markov process is characterized by the probabilities:

$$\text{Prob}\{\text{success following success}\} = 0.977$$
$$\text{Prob}\{\text{error following success}\} = 0.023$$
$$\text{Prob}\{\text{success following error}\} = 0.665$$
$$\text{Prob}\{\text{error following error}\} = 0.335$$

From these probabilities follows:

$$H_1(\text{loc}) = -0.034 \times (0.335 \log_2 0.335 + 0.665 \log_2 0.665) - 0.966 \times (0.023 \log_2 0.023 + 0.977 \log_2 0.977) = 0.185 \text{ bits/key}$$

Thus, this approximation reduces the error-location uncertainty by about one-sixth. Additional refinements are suggested by the data, but will not result in any significant reduction of uncertainty; e.g., introduction of error probabilities conditional upon the two preceding acts, gives a value $H_2(\text{loc})$ which is 0.185 bits/key, thus, not smaller than the value for $H_1(\text{loc})$.

The amount of information needed to correct errors depends equally on the error pattern. A given output, even if known to be erroneous, might still suggest a limited range of possible inputs. For instance, in playing piano, an error is likely to be not very far from the target key. We have used this consideration to bracket $H_{out}(\text{in})$ between values obtained by assuming error ranges which are certainly too large and too small, respectively. If no regularities in the error pattern are known, then all one knows about the input is that it is not the output; in the special case, frequently occurring in the laboratory, where one has k equiprobable input categories, and each input is equally likely to produce a faulty output, we have

$$H(\text{cor}) = g \log_2 (k-1)$$

In the binary case, H(cor) vanishes.

The method here discussed is useful in two ways. First, it establishes the relations between error rate, error pattern, and information transmission in a fashion which is lucid and easy to survey; thus, we found it convenient to use for a first computation of T in situations not previously dealt with. Second, the method enables one to perform an estimation of T in successive steps; it establishes upper bounds for the equivocation which together with the lower bound (zero) define an interval within which T must lie; any knowledge about error pattern can then be used to narrow down the interval.

Reference:

Shannon, C. and W. Weaver The Mathematical Theory of Communication. University of Illinois Press, Urbana, 1949.

UPPER BOUNDS FOR THE EQUIVOCATION

A. A. Blank

Shannon defines information transmitted, T(in,out), as

$$T(in,out) = H(in) - H_{out}(in)$$

where H(in) is the uncertainty of the source and $H_{out}(in)$ is the equivocation, or the uncertainty as to the source at the receiving end of a channel. Set $p_{ij} = P\{i,j\}$, the joint probability that the i-th input category is coupled with the j-th output category. We set

$$p_i = \sum_j p_{ij}; \qquad q_j = \sum_i p_{ij}$$

$$p(j|i) = \frac{p_{ij}}{p_i} \qquad q(i|j) = \frac{p_{ij}}{q_j}.$$

$$H(in) = -\sum p_i \log p_i$$

$$H_{out}(in) = -\sum_j q_j \sum_i q(i|j) \log q(i|j).$$

The quantity $H_j(in) = \sum_i q(i|j) \log q(i|j)$ is called the equivocation when the j-th output occurs or the uncertainty as to the source with respect to the j-th output. If the purpose of the channel is to duplicate the input, that is, to couple an input in the i-th category with an output in the same category, it is desirable to locate and correct errors. The quantity $H_j(in)$ may be represented as a sum

$$H_j(in) = H_j(loc) + H_j(cor)$$

2 Upper Bounds

Here

$$H_j(\text{loc}) = - q(j|j) \log q(j|j) - \left[1-q(j|j)\right] \log \left[1-q(j|j)\right].$$

$H_j(\text{loc})$ is called the error-locating information with respect to the j-th output and represents the conditional uncertainty of a source which reports on the truth or falsehood of an output in the j-th category. The quantity, $H_j(\text{cor})$ is called the correction information with respect to j and satisfies

$$H_j(\text{cor}) = - \left[1-q(j|j)\right] \sum_{(i\neq j)} \frac{q(i|j)}{1-q(j|j)} \log \frac{q(i|j)}{1-q(j|j)}$$

$$= - \sum_{i\neq j} q(i|j) \log q(i|j) + \left[1-q(j|j)\right] \log \left[1-q(j|j)\right]$$

$H_j(\text{cor})$ is the information required to correct an output on the condition that it falls in the j-th category.

As the error-locating information associated with the channel we take

$$H(\text{loc}) = - p \log p - (1-p) \log (1-p)$$

where $p = \sum_i p_{ii}$. As the correction information associated with the channel we take

$$H(\text{cor}) = - (1-p) \sum_i \frac{p_i - p_{ii}}{1-p} \log \frac{p_i - p_{ii}}{1-p}$$

Inequality 1: $H_{out}(in) \leq H(loc) + H(cor)$

Proof:

$$H_{out}(in) = -\sum_j q_j \sum_i q(i|j) \log q(i|j)$$

$$= -\sum_i \sum_j q_j\, q(i|j) \log q\,(i|j)$$

$$= -\sum_i q_i\, q(i|i) \log q(i|i)$$

$$-\sum_i \left\{ \sum_{j \neq i} q_j\, q(i|j) \log q(i|j) + q_i . 0 \log 0 \right\}$$

$$\leq -\left(\sum_i p_{ii} \right) \log \sum_i (p_{ii})$$

$$-\sum_i \left\{ \left(\sum_{j \neq i} p_{ij} \right) \log \left(\sum_{j \neq i} p_{ij} \right) \right\}$$

$$\leq -\, p \log p - \sum_i (p_i - p_{ii}) \log (p_i - p_{ii}).$$

Since the last term on the right is equal to H(loc) + H(cor) the proof is complete. Inequality 1 may be replaced by stronger inequalities with respect to H(loc) and H(cor). In particular, it will be shown that $\sum q_j\, H_j\,(loc) \leq H(loc)$ and $\sum q_j\, H_j\,(cor) \leq H(cor)$.

4 Upper Bounds

Inequality 2: $\sum q_j \, H_j(\text{loc}) \leq H(\text{loc})$

Proof:

$$\sum q_j \, H_j(\text{loc}) = - \sum_j q_j \, q(j|j) \log q(j|j)$$

$$- \sum_j q_j \left[1-q(j|j)\right] \log \left[1-q(j|j)\right]$$

$$\leq - \sum_j q_j \, q(j|j) \log \sum_j q_j \, q(j|j)$$

$$- \sum_j q_j \left[1-q(j|j)\right] \log \sum_j q_j \left[1-q(j|j)\right]$$

$$\leq - \sum_j p_{jj} \log \sum p_{jj}$$

$$- \sum_j (q_j - p_{jj}) \log \sum_j (q_j - p_{jj})$$

$$\leq - p \log p - (1-p) \log (1-p) = H(\text{loc}).$$

Inequality 3: $\sum q_j H_j(\text{cor}) \leq H(\text{cor})$.

The proof of this inequality requires a preliminary result.

Lemma: $$-\sum \lambda_i \log \frac{\sum \lambda_j}{\sum \sigma_j} \geq -\sum \lambda_i \log \frac{\lambda_i}{\sigma_i}$$

provided only that $\sigma_i, \lambda_i \geq 0$, $\sum \lambda_i \leq 1$.

Proof of Lemma:

$$\sum \lambda_i \log \frac{\sum \lambda_i}{\sum \sigma_j} \cdot \frac{\sigma_i}{\lambda_i} = \sum \lambda_i \log \frac{\sum \lambda_j}{\sum \sigma_j} \cdot \frac{\sigma_i}{\lambda_i} + (1-\sum \lambda_j) \log 1$$

$$\leq \log \left[1-\sum \lambda_j + \sum_i \frac{\sum \lambda_j}{\sum \sigma_j} \sigma_i\right]$$

$$\leq \log \left[1- \sum \lambda_j + \frac{\sum \lambda_j}{\sum \sigma_j} \sum_i \sigma_i\right] = \log 1$$

$$\leq 0.$$

Since $$\sum \lambda_i \left(\log \frac{\sum \lambda_i}{\sum \sigma_j} - \log \frac{\lambda_i}{\sigma_i}\right) \leq 0$$

we have

$$-\sum \lambda_i \log \frac{\lambda_i}{\sigma_i} \leq -\sum \lambda_i \log \frac{\sum \lambda_i}{\sum \sigma_j}$$

Proof of Inequality 3:

$$-\sum_j q_j H_j(cor) = \sum_{\substack{i \\ (i\neq j)}} \sum_j p_{ij} \log \frac{p_{ij}}{q_j - p_{jj}}$$

$$= -\sum_i \left\{ \sum_{\substack{j \\ j\neq i}} p_{ij} \log \frac{p_{ij}}{q_j - p_{jj}} \right\}$$

Noting that $\sum_{\substack{j \\ (j\neq i)}} p_{ij} \leq 1$,

we have by the lemma above,

$$-\sum_j q_j H_j(cor) \leq -\sum_i \left\{ \sum_{j\neq i} p_{ij} \log \frac{\sum_{j\neq i} p_{ij}}{\sum_{j\neq i} q_j - p_{jj}} \right\}$$

$$\leq -\sum_i \left\{ (p_i - p_{ii}) \log \frac{(p_i - p_{ii})}{(1-p - (q_i - p_{ii})} \right\}$$

$$\leq -\sum_i (p_i - p_{ii}) \log (p_i - p_{ii}) + \sum_i (p_i - p_{ii}) \log \left[(1-p) - (q_i - p_{ii}) \right]$$

$$\leq -\sum_i (p_i - p_{ii}) \log (p_i - p_{ii}) + \sum_i (p_i - p_{ii}) \log (1-p)$$

$$\leq -\sum_i (p_i - p_{ii}) \log (p_i - p_{ii}) + (1-p) \log (1-p)$$

$$\leq H(loc).$$

The value $H_{out}(in)$ is over-estimated by the value $\tilde{H}_{out}(in)$ = H(loc) + H(cor). It follows that

$$H(in) - \tilde{H}_{out}(in) \leq T(in;out)$$

In other words, we err on the conservative side if we take $\tilde{H}_{out}(in)$ as an estimate of $H_{out}(in)$. The final result may be considered as an example of the general statement that any source which fully corrects the errors of transmission must have an entropy no less than $H_{out}(in)$.

REMARKS ABOUT
THE METHOD OF HINTS

H. Quastler

A subject receives a message. One wishes to establish how much of the message was assimilated. "How much" is taken to mean "how many information units" (or "bits"). If the S is able to transmit the message correctly, then he certainly has assimilated its information content. If his transmission is only partially correct, then we may give the S hints containing some auxiliary information which will help him to correct his errors and reconstruct the entire message. It can be shown that

$$\left[\begin{array}{c}\text{Amount of information} \\ \text{assimilated}\end{array}\right] \geqslant \left[\begin{array}{c}\text{Amount of infor-} \\ \text{mation in input}\end{array}\right] - \left[\begin{array}{c}\text{Amount of} \\ \text{information} \\ \text{in hints}\end{array}\right]$$

where all amounts of information are measured in the same units (bits).

There are many ways of giving hints. A good method will be one which fulfills the following conditions:

(i) it permits accurate estimation of the amount of information contained in the hints,

(ii) it does not confuse the S,

(iii) it gives as little information as possible.

The issue can be made clear by reference to a familiar situation. A student is being examined; one wishes to give him credit for what he knows. Notoriously, his first answers

do not reveal the full extent of his knowledge. The examiner tries to help with hints. He must try to keep the actual information in the hints small (otherwise the student could not be given credit for the answer) and not to confuse the student. The requirement dealing with numerical estimation does not apply, since no quantitative estimation of the knowledge is obtained.

The method of hints is analogous to an often-used method of estimating amounts of information transmitted:

$$\left[\begin{array}{c}\text{Amount of}\\ \text{information}\\ \text{transmitted}\end{array}\right] \geqslant \left[\begin{array}{c}\text{Amount of}\\ \text{information}\\ \text{in input}\end{array}\right] - \left[\begin{array}{c}\text{Amount of information}\\ \text{needed to locate}\\ \text{and correct errors}\end{array}\right]$$

But, the two are not identical. It is true that the hints supply "information needed to locate and correct errors"; however, in addition to this information S may use information stored in his memory but not utilized in his first statement.

For instance, a situation like the following might happen: the display is a dot on a vertical line; the line is thought to be divided into intervals of equal length; S is asked to state which interval contains the dot. Suppose he makes errors with an over-all probability q; and that, in case of error, he is too high by one interval with a probability of α, too low by one interval with a probability of $(1-\alpha)$; no other errors occur. Then

$$\left[\begin{array}{c}\text{Amount of information}\\ \text{needed to locate errors,}\\ \text{per act}\end{array}\right] = -q \log_2 q - (1-q) \log_2 (1-q)$$

$$\left[\begin{array}{l}\text{Amount of information} \\ \text{needed to} \\ \text{correct errors,} \\ \text{per act}\end{array}\right] = -q\left[\alpha \log_2 \alpha + (1-\alpha)\log_2 (1-\alpha)\right]$$

These two quantities must be deducted from the input information to obtain the amount of information transmitted. Suppose, now, that S has some recollection about the direction of his error; then, telling him when he has committed an error will be all he needs to produce the correct statement. Thus, the amount of information assimilated will be greater than the amount of information transmitted (in the first statement) by the amount $\left\{-q\left[\alpha \log_2 \alpha + (1-\alpha)\, q\ \ (1-\alpha)\right]\right\}$. Some experiences, reported elsewhere in this collection, indicate that the example given, while greatly simplified, is not unrealistic.

In general, the estimated "amount of information assimilated" may be larger or smaller than the estimated "amount of information transmitted," depending on the amount of such retained information, on the efficiency of giving and utilizing hints, and on the efficiency of the error locating and correcting code. Ordinarily, we do not expect the two to differ widely from each other.

The following paper, by A. A. Blank, gives a model of performance for one particular Method of Hints. It is the only model which has been worked up in some detail, and was used on experimental data.

A METHOD OF HINTS

A. A. Blank

Let us suppose we have a stochastic source of independent inputs a_i $(i = 1, \ldots, r)$, and a response generator which has the conditional probability $q(i|j)$ of a_i having been emitted by the source when the response is a_j and each response is independent of any other. The uncertainty as to the source when the symbol a_j has been reported is defined as

$$H(in) = -\sum_i q(i|j) \log q(i|j).$$

Suppose again that when the receiver is in error the fact of the occurrence of error is registered and fed back and the response generator is then constrained to report differently. We shall consider two possibilities:

1. Complete utilization of the order of probability.

Let the conditional probabilities of error be ordered

$$q(i_2|j) \geqslant q(i_3|j) \geqslant \cdots \geqslant q(i_r|j)$$

where $i_2, \ldots, i_r$ is some permutation of the indices excluding j. We shall suppose that, in the event of error, the response generator will report symbols in decreasing order of probability until the correct one is reached. Let $Q_\nu(j)$ denote the frequency with which j is reported correctly at the ν-th report. We have

$$Q_\nu(j) = a(i_\nu|j) \qquad (\nu = 1, \ldots, r)$$

where we set $i_1 = j$. Clearly

$$H_j(in) = - \sum_{\nu} Q_\nu(j) \log Q_\nu(j).$$

The equivocation or uncertainty as to the source is given by

$$H_{out}(in) = - \sum_{j} q_j H_j(x)$$

$$= - \sum_{j} \sum_{\nu} q_j Q_\nu(j) \log Q_\nu(j)$$

Hence, with complete knowledge of the order of probability of the various errors we may compute the equivocation by using the method of hints (feed back an error message) and tabulating the probability for each initial response of obtaining the correct report at the ν-th stage.

2. Incomplete utilization of the order of probability.

Let us suppose that the generator reports in the order of decreasing probability until the k-th stage and that from the (k + 1)-th stage on responses are chosen with equal probability from the remaining alternatives. In that case set

$$Q_\nu(j) = q(i_\nu | j) \qquad \nu = (1 \ldots k)$$

$$P(j) = 1 - \sum_{1}^{k} Q_\nu(j)$$

We have,

$$H_j^*(in) = - \sum_{\nu=1}^{k} Q_\nu(j) \log Q_\nu(j) - P(j) \log \frac{P(j)}{r-k} \geqslant H_j(in)$$

$H_j^*(in)$ represents the maximum possible equivocation with respect to the source if the order of the first k conditional probabilities of response are utilized in the prescribed manner. This estimate sacrifices only knowledge with respect to the rarer events.

The inequality above leads to the value $\sum q_j H_j^*(x)$ as an upper estimate for $H_{out}(in)$,

$$\sum q_j H_j(in) \geqslant H_{out}(in)$$

The computation of this estimate requires less than the computation of $H_{out}(in)$ since it requires $\bar{k}\, r$ categories ($\bar{k} = \frac{1}{r} \times \sum_j k(j) < r$), instead of r^2 categories.

It is not difficult to establish, with the same data, lower estimates for $H_j(in)$ and hence $H_{out}(in)$. We set

$$H_j'(in) = - \sum_{\nu=1}^{k} Q_\nu(j) \log Q_\nu(j) - P(j) \log P(j)$$

where $P(j) = 1 - \sum_{\nu=1}^{k} Q_\nu(j) = \sum_{\nu=k+1}^{r} q(i_\nu|j)$

$H_j'(in)$ represents the minimum equivocation with respect to the source if the first k responses are ordered according to the conditional probabilities of initial response. From the inequality

$$- \sum_{\nu=k+1}^{r} q(i_\nu|j) \log q(i_\nu|j) \geqslant - \sum_{\nu=k+1}^{r} q(i_\nu|j) \log \sum_{\nu=k+1}^{r} q(i_\nu|j)$$

it follows that

$$H_j'(in) \leqslant H_j(in)$$

4 A Method of Hints

A more refined estimate is sometimes given by

$$H_j''(in) = - \sum_{\nu=1}^{k} Q_\nu(j) \log Q_\nu(j) - P(j) \log q_k(j)$$

$H_j''(in)$ approximates the minimum equivocation with respect to the source if the first k responses are ordered according to the conditional probabilities of initial response and if the probability of correct response at any stage does not exceed that at any preceding stage. From the inequality $q(i_k|j) = Q_k(j) \geqslant q(i_\nu|j)$ for $\nu > k$, we have

$$- \sum_{\nu=k+1}^{r} q(i_\nu|j) \log q(i_\nu|j) \geqslant - \sum_{\nu=k+1}^{r} q(i_\nu|j) \log q(i_k|j)$$

and hence,

$$\overline{H}_j''(in) \leqslant H_j(in)$$

If $Q_k(j) < P(j)$ then $H_j''(in)$ is a better estimate of $H_j(in)$ than $H_j'(in)$. Let

$$\overline{H}_j(in) = \text{Max}\left(H_j'(in), H_j''(in)\right) \leqslant H_j(in).$$

where $\overline{H}_j(in)$ is the better approximation to $H_j(in)$ of the two lower bounds $H_j'(in)$ and $H_j''(in)$. We have

$$\sum q_j \overline{H}_j(in) \leqslant H_{out}(in) \leqslant \sum q_j H_j^*(in).$$

3. Pooled data.

In some instances it may be impossible to obtain reliable values of the frequency of correct response at the ν-th stage

for each initial response. One may only know or be able to utilize the frequency of correct response at the ν-th stage taken over all initial reports. This would be the value

$$Q_\nu = \sum_j q_j \, Q_\nu(j).$$

We may estimate $H_{out}(in)$ by

$$\tilde{H}_{out}(in) = - \sum_{\nu=1}^{r} Q_\nu \log Q_\nu \geqslant - \sum_j q_j H_j(in) \geqslant H_{out}(in).$$

If no data are obtained beyond the k-th response we may be sure that $\tilde{H}_{out}(in)$ is bounded above and below by

$$\tilde{\tilde{H}}_{out}(in) \leqslant \tilde{H}_{out}(in) \leqslant \tilde{H}^*_{out}(in)$$

where

$$\tilde{H}^*_{out}(in) = - \sum_{\nu=1}^{k} Q_\nu \log Q_\nu - P \log \frac{P}{r-k}$$

and

$$\tilde{\tilde{H}}_{out}(in) = \text{Max} \left(\tilde{H}'_{out}(in), \tilde{H}''_{out}(in) \right) .$$

Here

$$P = 1 - \sum_{\nu=1}^{k} Q_\nu$$

and

$$H'_{out}(in) = - \sum_{\nu=1}^{k} Q_\nu \log Q_\nu - P \log P$$

$$H''_{out}(in) = - \sum_{\nu=1}^{k} Q_\nu \log Q_\nu - P \log Q_k .$$

The barred symbols represent a lumping of response data into fewer categories and the starred symbols represent the equal

division of response data into available categories.

We cannot be sure of the relationship of $\tilde{\tilde{H}}_{out}(in)$ to $H_{out}(in)$, but we do know that $\tilde{H}^{*}_{out}(in)$ is an overestimate.

To obtain a lower estimate of $H_{out}(in)$ from pooled data we may proceed as follows:

We have $$H_{out}(in) = H(in,out) - H(out)$$

Now $$H(in,out) = - \sum_{\nu=1}^{r} \sum_{j=1}^{r} p_{\nu j} \log p_{\nu j}$$

where $$p_{\nu j} = q_j \, Q_\nu(j).$$

Clearly $$\Sigma_j \; p_{\nu j} = Q_\nu.$$

Now $$- \sum_{j=1}^{r} p_{\nu j} \log p_{\nu j} \geqslant - \sum_{j=1}^{r} p_{\nu j} \log \sum_{j=1}^{r} p_{\nu j} = - Q_\nu \log Q_\nu$$

Hence, $$H_{out}(in) \geqslant - \Sigma \, Q_\nu \log Q_\nu - H(out)$$

or $$H_{out}(in) \geqslant - \sum_{\nu=1}^{r} Q_\nu \log Q_\nu + \sum_{j=1}^{r} q_j \log q_j = \tilde{H}_{out}(in) - H(out).$$

If we omit the data for the categories $\nu = k+1,\ldots,r$, we obtain a still lower estimate

$$H_{out}(in) \geqslant \tilde{H}_{out}(in) - H(out) \geqslant \tilde{\tilde{H}}_{out}(in) - H(out)$$

These last estimates are very crude. Zero should often be a better lower estimate of $H_{out}(in)$.

THE UNCERTAINTY MEASURE FOR QUANTIZED NORMAL DISTRIBUTIONS

A. A. Blank

Whenever a discrete random variable may be thought of as having its values imbedded in a continuum there is a suggestion that it may be convenient to consider the distribution of the discrete variable as the quantization of the distribution of some continuous random variable. In the usual game of darts, for example, the compartment in which the dart sticks is a discrete variable, but represents in an obvious way the quantization of the error distribution in attempting a strike at the bull's eye.

In some communication problems the discrete outcomes arising may also be thought of as a quantization of a continuous distribution (e.g., when a dial is read to the nearest tenth of a division). In information theory, the uncertainty function is not defined in the same way for discrete and continuous distributions. The uncertainty for a discrete variable is defined as

$$(1) \qquad H = -\sum p_k \log p_k$$

where the value p_k denotes the probability that the variable will take its k-th value. In the continuous case, the uncertainty is defined by the integral

$$(2) \qquad H' = -\int_{\Omega} \log f(\xi)\, d F(\xi)$$

where ξ denotes the random variable, $f(\xi)$ is its probabil-

ity density, $F(\xi)$ its cumulative distribution function and Ω is the sample space in which ξ varies. As immediate points of difference between the two definitions it will be observed that H is non-negative while H' may assume any real value whatever.

The most useful distribution to treat in this manner is the quantized normal distribution. The normal density function

$$(3) \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

gives

$$(4) \qquad H' = \frac{1}{2} \log 2\pi e \sigma^2$$

and since σ^2 may be any non-negative value H' may assume any value from minus to plus infinity. If a discrete distribution may be interpreted as a quantization of a normal distribution it is especially convenient to use (4) since all we have to do to compute H' is to find σ. H' will be a good approximato H if the distribution is quantized into units sufficiently small with respect to σ. If the precision is high and most of the values of the discrete variable fall in one class, then H will be very close to zero but H' will take on large negative values; in this case, therefore, the approximation cannot be used.

Let us see generally how H' is related to H. If the axis of the continuous variable is broken into intervals of

size Δ *, and $\lambda = \frac{\sigma}{\Delta}$, then H may be computed as a function of λ from (1) where

$$p_k(\lambda) = \int_{(k - \frac{1}{2})\lambda}^{(k + \frac{1}{2})\lambda} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\, dt \quad (k = 0, \pm 1, \pm 2, \ldots)$$

The values of $H'(\lambda)$ and $H(\lambda)$ are both plotted against λ on the accompanying graph. For $\lambda > 2$, that is, $\Delta < \frac{\sigma}{2}$ we may for all practical purposes assume $H' = H$. For $\lambda < 0.1$, or $\Delta > 10\sigma$, we may assume $H = 0$. A nomogram for intermediate values of λ is included.

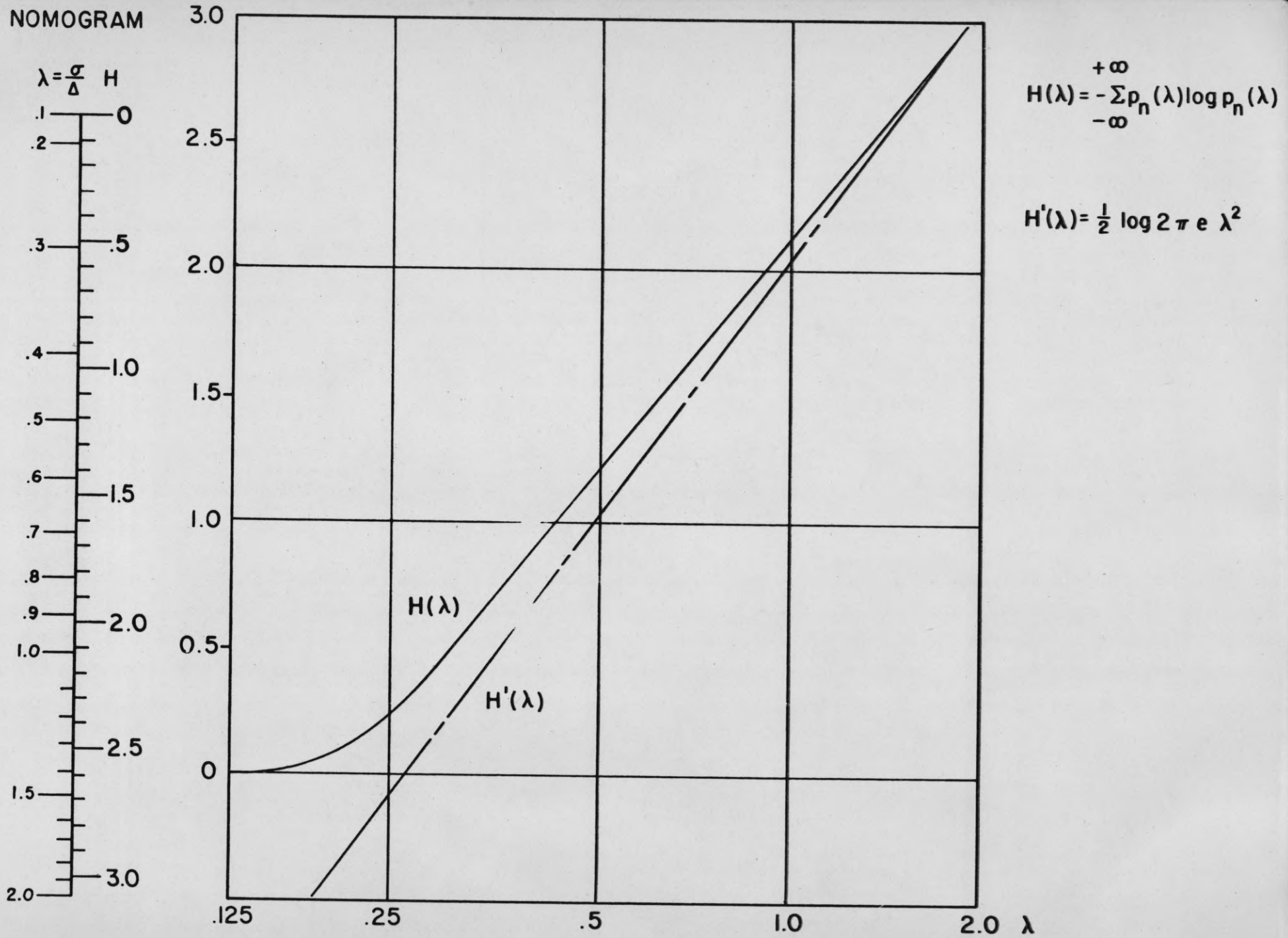
In effect, when a distribution is obtained which may be assumed to be quantized normal, the associated uncertainty may be estimated by obtaining the variance of the corresponding continuous normal distribution and looking up $H(\lambda)$ on the graph. Alternatively, one may obtain $H'(\lambda)$ and look up $H(\lambda)$ on the nomogram.

---

* Sheppard's estimate then gives σ by

$$\sigma^2 = \sigma_d^2 - \frac{1}{12}\Delta^2$$

where σ_d is the variance of the quantized distribution.


NOMOGRAM
$\lambda = \frac{\sigma}{\Delta}$ H
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
1.5
2.0
0
.5
1.0
1.5
2.0
2.5
3.0
3.0
2.5
2.0
1.5
1.0
0.5
0
.125
.25
.5
1.0
2.0 λ
H(λ)
H'(λ)
$H(\lambda) = -\sum_{-\infty}^{+\infty} p_n(\lambda) \log p_n(\lambda)$
$H'(\lambda) = \frac{1}{2} \log 2\pi e \lambda^2$

DIFFERENT METHODS COMPARED

H. Quastler.

A - "INFORMATION TRANSMITTED" VS. "PERCENT SUCCESS"

Among the measures which can be used assess human performances are "information transmitted" and "percent success." The two measures are often roughly proportional.

The amount of information transmitted can be thought of as the sum of information transmitted in successful acts, minus information needed to locate successful acts, plus information transmitted in errors (due to "near misses" and other regularities of error patterns). Using the symbols:

p . . . per cent success

H(in) . . . Information input, per act

H(loc) . . . Information needed to locate errors, per act

H(cor) . . . Information needed to correct errors, per act

T(in;out) . . Information transmitted, per act

we have

$$\left[\begin{matrix}\text{Information transmitted} \\ \text{in successful acts}\end{matrix}\right] = p \cdot H(in)$$

$$\left[\begin{matrix}\text{Information transmitted} \\ \text{in errors}\end{matrix}\right] = (1-p) \cdot H(in) - H(cor)$$

and

$$T(in;out) = H(in)\left[p+1-p\right] - H(loc) - H(cor)$$

The claim made is that often

$$p \cdot H(in) \approx T(in;out).$$

Now, this is certainly not generally true. In fact, one can have perfect transmission with nothing but errors (e.g., if a transducer receives a binary message and passes it on after inverting each symbol); it is also possible that no information at all is carried in errors, in which case p . H(in) is an overestimate of T(in;out). Ordinarily, neither extreme is likely to occur. For moderate error rates, and a moderate amount of information carried in errors, the approximation will not be too bad. This is shown graphically in the figure.

The following two extreme examples are taken from work done in this laboratory. In one (A. A. Blank), S had to recognize letters; H(in) per letter was varied by using various constraints. In this case, there is very little information carried in errors; in most cases, S recognizes a letter accurately or not at all. Hence, p . H(in) is larger than T(in;out):

Input:	H(in), per letter	Sub. A		Sub. B	
		p.H(in)	T(in;out)*	p.H(in)	T(in;out)*
Single equiprobable letters	4.7	2.2	1.5	2.5	1.9
Letters, English frequencies	4.1	2.3	1.4	2.5	1.7
Pairs of initial letters	3.2	2.3	1.4	1.9	0.9
4-letter words	1.8	1.6	1.0	1.7	1.5

* estimated by Method of Hints

The next example (J. W. Osborne and K. S. Tweedell) deals with the task of locating a marker on a scale. In this case, near-misses are the rule. Thus, errors carry considerable information, and p . H(in) is sometimes less than T(in;out):

no. of intervals in scale	H(in)	Sub. F p.H(in)	Sub. F T(in;out)	Sub. W p.H(in)	Sub. W T(in;out)
16	4.0	3.3	3.2	4.0	4.0
20	4.3	3.2	3.2	-	-
24	4.6	2.3	2.9	4.2	4.1
32	5.0	2.7	3.1	3.5	3.8
36	5.2	2.9	3.3	3.6	3.7
48	5.5	2.5	3.2	4.5	4.3

B - "ERROR MAGNITUDE" VS. "HINTS"

This comparison was made during an investigation of the amount of information assimilated from a single-pulse display. The display contained one or two strips or dials, divided into discrete intervals; a marker was placed at the center of any one interval; the subject had to state in which interval it was (J.W. Osborne and K.S. Tweedell).

In four particular runs, the equivocation was evaluated both by the method of hints and by error magnitudes. In the former case, the procedure followed was that described by A. A. Blank ("A Method of Hints," this collection); in the latter case, the procedure was based on the discrepancy between the input and S's first statement. Let $\underline{\mathbb{Z}}$ be the magnitude of this error, in scale intervals; ν_i be the observed frequency of errors of magnutude i; and $H(\mathbb{Z})$ the uncertainty concerning the error magnitude; then

$$H(\mathbb{Z}) = - \sum_i \frac{\nu_i}{\nu} \log_2 \frac{\nu_i}{\nu} \qquad \left(i = 0, +1, -1, +2, \ldots. \quad \sum_i \nu_i = \nu \right)$$

4 Different Methods

If H(Z) would be evaluated separately for each output category, then it would be an accurate estimate of the equivocation; if all output categories are lumped (as in this case) it is a lower bound.

The following table shows the results:

Display	no. of tests	trials per test	mean H(in)	mean est. T(in;out)	
				error magnitude	hints
horizontal strip with intervals in black & white	5	80-120	4.77	3.77 [2]	3.90
horizontal strip (intervals blank)	5	80-120	4.77	3.47 [2]	3.52
2 strips, 24 [1] intervals, marked black and white	8	80	4.58	2.51 [3]	2.62
2 dials, 24 [1] intervals	8	80	4.58	3.44	3.52

1) results given for single strip or dial.
2) terminal intervals treated separately.
3) no separate treatment of terminal intervals.

In all four cases, the estimation by the method of hints gave slightly higher values for T(in;out). This might mean that S has assimilated some information which does not appear in his first response, but is produced in response to hints.

C - "ERROR MAGNITUDE" VS. "VARIANCE"

This comparison was based on data obtained in a recognition experiment. The display was a dot in a square. The dot could assume a limited number of positions in the square.

The subject was given a score sheet on which the permissible positions were marked; he had to state in which of these positions the dot was located. (J. W. Osborne & K. S. Tweedell).

We evaluated, separately, location vertically and horizontally. The data were worked up by evaluation of error magnitude, as described above. In addition, we computed the variance, or mean squared deviation between input and output, lumping all output categories. From the variance, the equivocation can be computed (see A. A. Blank, "The Uncertainty Measure for Quantized Normal Distributions", this volume).

The following table shows the amounts of equivocation, by the two alternate methods; each entry is based on 40 trials:

No. of possible positions:	Method of computing equivocation	Subject			
		1	2	3	4
15 x 15	E.M.[1]	1.8	2.2	1.6	1.9
	V.[2]	2.0	2.4	1.8	1.8
19 x 19	E.M.	2.2	2.0	1.9	2.0
	V.	2.2	2.2	2.0	1.9
31 x 31	E.M.	2.9	3.2	2.7	2.7
	V.	3.2	3.2	2.7	3.0

1) Error magnitude.
2) Variance.

One sees that the "Variance" method tends to give slightly higher values for the equivocation. This is caused by the presence of a few very large errors, which contribute greatly to the variance, but not much to the uncertainty measure.

Straight Line: $q \cdot H(in)$ = Amount of Information Not Transmitted

Curved Line: Maximum Equivocation = $-q \log_2 q - (1-q) \log_2 (1-q) + q \log_2 (k-1)$

Hatched Area: Range of Equivocation

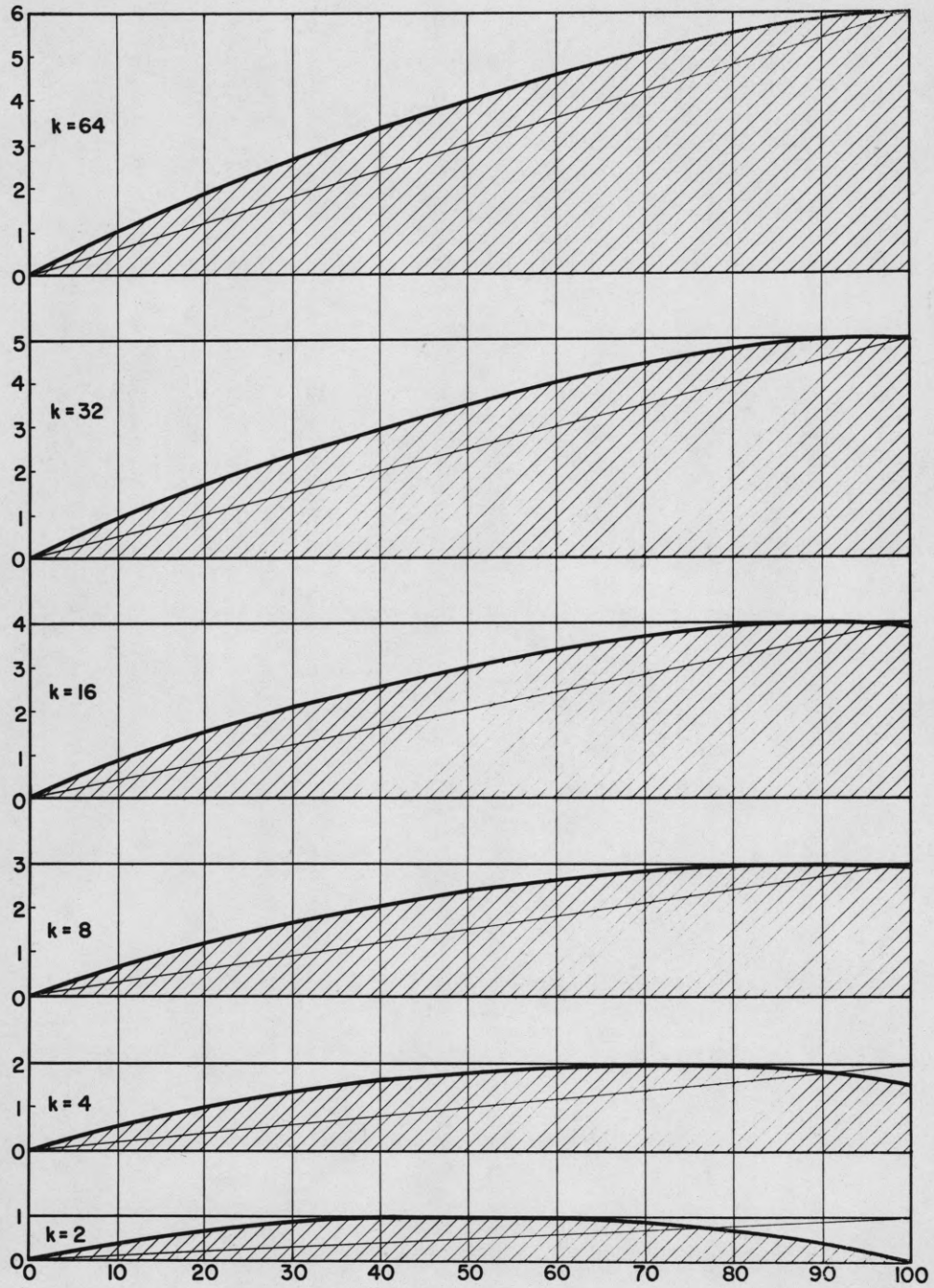


FIG. EQUIVOCATION VS. INFORMATION NOT TRANSMITTED

Abscissa: Percentage of Error (q)
Ordinate: H_{OUT} (in) in Bits
k = No. of Equiprobable Input Categories