CHAPTER 4

INFORMATION OF EVENTS WITH DISCRETE OUTCOMES: METHODS OF COMMUNICATION THEORY AND PSYCHOLOGY

We were introduced to the concept of *information* in Chapter 2, where the distinction was made (note 4) between discrete and continuous outcomes to an event. An event with only discrete outcomes is one such as the toss of a coin, which can land only heads or tails. An event with continuous outcomes is one such as the time of a response, which can, in principle, take on an infinity of values. In this chapter we consider only events of the former type, and we analyze, using two parallel and equivalent methods, how information is transferred during these events. The reader is reminded that the word *information*, if it is not qualified by an adjective, will always refer to that quantity which is measured by information theoretical entropy (sometimes called *communications entropy*). In Chapter 2 we glimpsed the way in which the information or entropy concept was to be employed in the analysis of sensory events: a natural law will be formulated giving perceptual response (such as the rate of action potential propagation in a sensory neuron) as a mathematical function of the entropy of a stimulus. If information, and hence entropy, were only an arbitrary creation of communications engineers, fabricated as a convenient means of measuring the efficiency of transmitting a message in a telephone cable, it would be surprising indeed that nature would use the same measure in sensory neurons. Therefore, we shall proceed in Chapter 6 to examine how the information concept, masquerading in different garb, was introduced into physics by Ludwig Boltzmann more than half a century before Shannon's paper on the theory of communication, and how information (or, equivalently, informational entropy) was woven into the fabric of physical law. As the story unfolds, we shall see that the entrance of informational entropy as a primary variable of neurophysiology seems to be an extension of its role as a primary variable in physics.

PICKING UP THE THREAD

We suppose that some event may happen in N discrete ways. For example, a election may result in only 1 winner from among 12 candidates who are running. We may say, therefore, that the election event has $N = 12$ possible discrete outcomes. The uncertainty that prevails before the election results are known is measured using the entropy, H, that was defined by Equation (2.1):

$$H = -\sum_{i=1}^{N} p_i \log p_i \tag{4.1}$$

where p_i is the probability of the i^{th} outcome. That is, H is a weighted sum of the logarithms of the *a priori* probabilities. Of course, the probabilities must sum to unity; that is,

$$\sum_{i=1}^{N} p_i = 1 \, . \tag{4.2}$$

Therefore, the election uncertainty is given by

$$H = -\sum_{i=1}^{12} p_i \log p_i \,,$$

where p_1 is the *assumed* probability that candidate 1 will be elected, etc. We note that, in contrast to the earlier example of tossing a coin or rolling a die where the *a priori* probabilities were possibly determined geometrically, the probabilities in this example are established by various subjective means (perhaps augmented by the results of pre-election polls). When the results of the election become known, the uncertainty vanishes, and information about the election takes its place. The information about which candidate won the election is equal to the preexisting uncertainty, so that

$$\mathscr{I} = H \,. \tag{4.3}$$

We recall that the base of the logarithms used is arbitrary; and also that when all the probabilities are equal (say equal to p), then

$$H = \log N \,, \tag{4.4}$$

where

$$N = 1/p \,. \tag{4.5}$$

We might ask the very natural question: Given some value for the number of possible outcomes, N, which values for p_i will render H maximum? For example, suppose we deal with an event with two possible outcomes whose probabilities are p_1 and p_2. Then, as a consequence of Equation (4.2), $p_2 = 1 - p_1$. What value of p_1 will then produce a maximum value for H? From Equation (4.1),

$$H = -p_1 \log p_1 - (1 - p_1)\log(1 - p_1) \,. \tag{4.6}$$

We shall take $\lim_{p_1 \to 0} p_1 \log p_1 = 0$. Then we may see that for $p_1 = 0$, $H = 0$, and for $p_1 = 1$, $H = 0$. This result is quite reasonable since, when $p_1 = 0$, $p_2 = 1$, the second outcome is a certainty and, therefore, uncertainty, H, vanishes. H is similarly equal to zero for $p_1 = 1$, when the first outcome

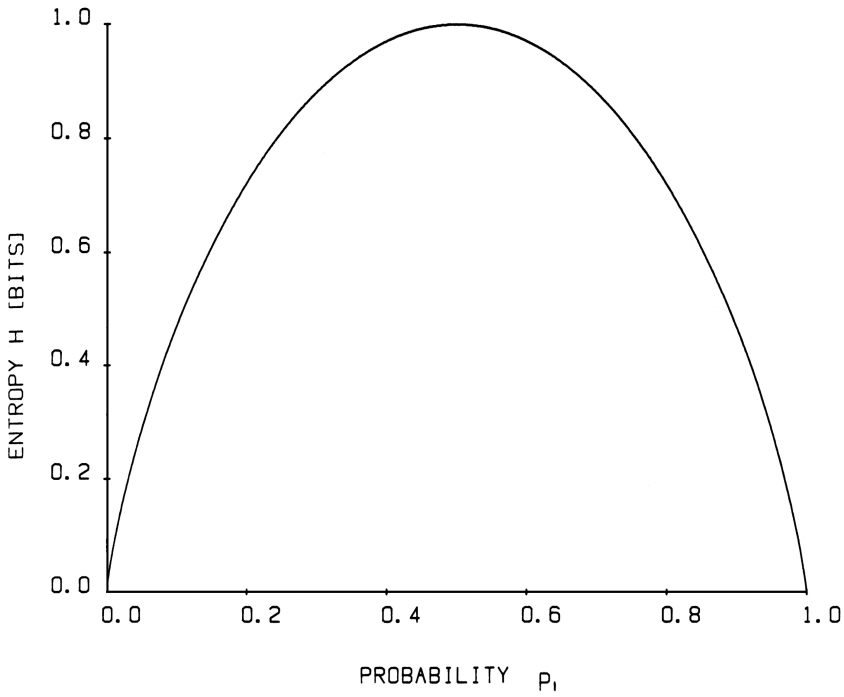

Figure 4.1 Entropy of an event with two possible outcomes, as a function of the probability, p_1, of one of these outcomes. $p_2 = 1 - p_1$. Note that entropy is maximum for $p_1 = p_2 = \frac{1}{2}$.

is a certainty. The mathematical function $H(p_1)$ must, therefore, be equal to zero for the two extreme values of p_1. Moreover, because of the symmetry in the variables p_1 and p_2, the function must be symmetrical about the line $p_1 = \frac{1}{2}$. The complete graph of H vs. p_1 is shown in Figure 4.1, where H is seen to be maximum for $p_1 = p_2 = \frac{1}{2}$.

Consider now the general case of an event with N possible outcomes. Then H is given by Equation (4.1) subject to the constraint expressed by Equation (4.2). In order to extremize H, that is, to find its relative maxima and minima, we introduce a Lagrangian multiplier, λ, to produce the expression

$$G = -\sum_{i=1}^{N} p_i \ln p_i + \lambda\left(\sum_{i=1}^{N} p_i - 1\right). \tag{4.7}$$

That is, we set up the expression

$$G = \text{entropy} + \text{Lagrangian multiplier} \times \text{constraint}.$$

The values of p_i for which H is an extremum subject to the normalization constraint is found by differentiating G partially with respect to each of the p_i, and equating the derivatives to zero:

$$\frac{\partial}{\partial p_k}\left[-\sum_{i=1}^{N} p_i \log p_i + \lambda\left(\sum_{i=1}^{N} p_i - 1\right)\right] = 0$$

$$-1 - \ln p_k + \lambda = 0$$

$$\ln p_k = \lambda - 1$$

$$p_k = e^{\lambda - 1}, \qquad \text{for all } p_k. \tag{4.8}$$

That is, all p_i are equal and must be equal to $1/N$ for an extremum.

In principle, we are not yet finished, since we must show that the extremum for H when $p_i = 1/N$ is, in fact, a maximum. For the completion of the proof, the reader is referred to Raisbeck (1963).

So the entropy, H, is maximum when the outcomes of the event are equally probable, as we saw in the example in Figure 4.1. In other words, we are most uncertain when an event is equally likely to occur in various possible ways, and we derive the greatest possible amount of information from

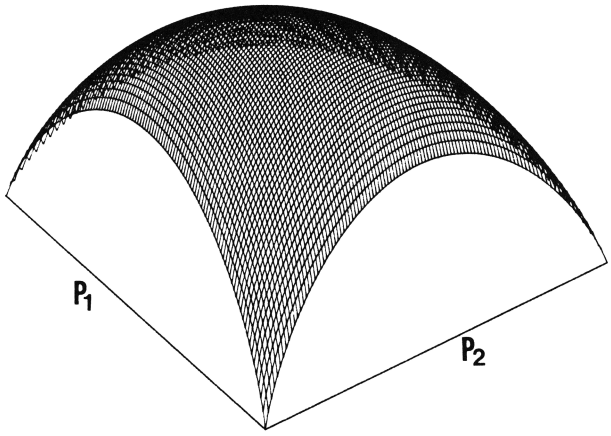

Figure 4.2 Entropy of an event with three possible outcomes, as a function of the probabilities p_1 and p_2 of these outcomes. Entropy, H, is given by $H = -\sum_{i=1}^{3} p_i \log p_i$. Since $p_3 = 1 - p_1 - p_2$, therefore

$$H = -p_1 \log p_1 - p_2 \log p_2 - (1 - p_1 - p_2)\log(1 - p_1 - p_2).$$

The graph shows H as a function of p_1 and p_2. However, appearances can be deceiving. In the graph shown, the origin is actually remote from the viewer and the p_1- and p_2- axes come toward him/her. The viewer's eye is situated below the p_1-p_2 plane and the surface is concave, not convex (as it appears). H is, of course, maximum at $p_1 = p_2 = p_3 = 1/3$.

observing the outcome of such an event. Just for the fun of it, Figure 4.2 depicts an event with 3 possible outcomes with probabilities p_1, p_2 and p_3. Since $p_3 = 1 - p_1 - p_2$, H can be plotted along the z-axis as a function of the two variables p_1 and p_2. H is maximum for $p_1 = p_2 = p_3 = 1/3$.

Just a note on the appropriateness of the log-function as a means of expressing information. Suppose we toss a fair coin. Then the head-tail information we obtain is $\log_2 2 = 1$ bit. If we toss the coin a second time, we receive an additional 1 bit of information. Therefore, the total amount of information that we receive by observing the results of 2 sequential tosses is 2 bits. Suppose, now, that we place 2 coins in a closed box, shake the box, and observe the results of the simultaneous toss. How much information do we receive? Well, there are 4 equally probable outcomes to the simultaneous toss: HH, HT, TH, TT. Therefore, the quantity of information received is equal to $\log_2 4 = 2$ bits, the same amount we received by tossing the two coins sequentially. Any other result would have been untenable.

A CHANNEL OF COMMUNICATION

We consider now some means of transmitting a message between two stations. The means is arbitrary. It could be electrical or optical, such as that used for a telephone, or even acoustical such as that used for ordinary speech. In order to simplify the initial discussion, suppose that only two symbols are transmitted: 1 or 0; that is, messages consist purely of strings of 0's and 1's. Let us suppose, also, that our channel transmits without error, so that each time the transmitter sends 0, the receiver gets 0, etc. For purposes of illustration, suppose that the probability of transmitting 0 is 0.2 and the probability of transmitting 1 is 0.8 (Figure 4.3). We can represent the probabilities of symbols (0, 1) for the transmitter (source) by the vector (0.2, 0.8), and the probabilities of symbols (0, 1) for the receiver by the same vector (0.2, 0.8). The transmitter, or *source entropy* is given by

$$H_{\text{source}} = -\sum_{i=1}^{2} p_s \log p_s = -0.2\log 0.2 - 0.8\log 0.8 .$$

Similarly, the *receiver entropy* is given by

$$H_{\text{receiver}} = -\sum_{i=1}^{2} p_r \log p_r = -0.2\log 0.2 - 0.8\log 0.8 .$$

The information received by the receiver is then

$$\mathcal{I} = H_{\text{source}} = H_{\text{receiver}} . \tag{4.9}$$

Now the realities of communication are such that interference, or noise, usually affects the transmission of signals through a channel, resulting in errors at the receiver. That is, the transmission of a zero will sometimes result in the receipt of a one and vice versa. In this case, the information received, $\mathcal{I}$, will not be equal to H_{source} or to H_{receiver}. We shall now develop equations governing the information received when a signal is received from such a noisy channel. Only a little in the way of basic mathematics is required: some facility with iterated summation operators such as $\sum_{j=1}^{n} \sum_{k=1}^{n}$, and

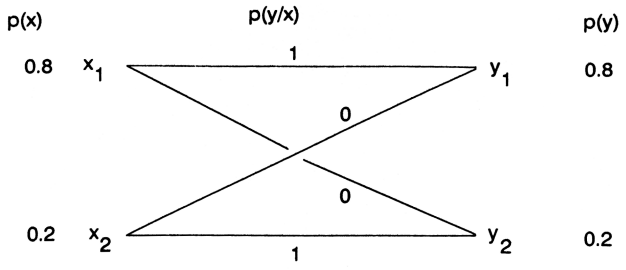

Figure 4.3 Information transmission in a noiseless channel. The probability of transmission of x_1 equals 0.8 and, since $p(y_1 \mid x_1) = 1$, therefore probability of receipt of y_1 equals 0.8, etc.

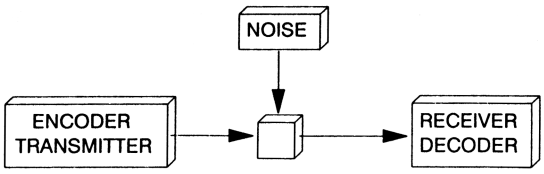

Figure 4.4 Schema for a noisy channel.

the definition of conditional probability. We deal with signals transmitted, for which we use the symbol x_j, and signals received, for which we shall use the symbol y_k. Let us generalize our lexicon of transmitted symbols from 2 to n. Then, for example, $p_j(x_1)$ represents the probability of transmitting signal x_1, and $p_k(y_3)$ represents the probability of receiving signal y_3. In general, then, $p_j(x_j)$ represents the probability of transmitting the signal x_j, and $p_k(y_k)$ represents the probability of receiving the signal y_k. To simplify the nomenclature, we can drop the subscripts following the p without introducing any ambiguity: $p(x_j)$ will mean $p_j(x_j)$, etc. We define $p(x_j|y_k)$ as the *conditional probability* of x_j given y_k; that is the probability that signal x_j was transmitted given that signal y_k was received. $p(y_k|x_j)$ is defined analogously. We further define $p(x_j, y_k)$ as the *joint probability* that x_j was transmitted and y_k received. From the definition of conditional probability emerge two fundamental equations that we use on various occasions:

$$p(x_j, y_k) = p(x_j|y_k) \cdot p(y_k) \tag{4.10}$$

$$p(x_j, y_k) = p(y_k|x_j) \cdot p(x_j) \tag{4.11}$$

where the dot signifies ordinary multiplication.

The sequence of events is now depicted by the well-known diagram shown in Figure 4.4.

A key feature of the noisy channel is what one might call *residual uncertainty*. The simple coin tossing paradigm involved, ostensibly, no noise, so that when we observed the outcome of a toss to be "heads," there were no lingering doubts that perhaps it was really "tails" and we had misread the face of the coin. For this reason, we could write simply

$$\mathcal{I} = H = -\sum p \ln p \, .$$

In the more general case, however, receipt of the signal y_k still leaves some residual uncertainty; $p(x_j|y_k)$ gives the probability, not always zero, that some signal other than x_j may have been transmitted. Therefore, $\mathcal{I}$ is not longer simply equal to $-\sum p(x_j) \ln p(x_j)$. In fact $\mathcal{I}$ is less than this amount due to the residual uncertainty. In general, for noisy channels

$$\mathcal{I} = H_{\text{before}} - H_{\text{after}} \, . \tag{4.12}$$

That is, the transmitted information is equal to the difference between the source entropy or uncertainty, H_{before}, given as usual by $-\sum p(x_j) \ln p(x_j)$, and the residual entropy, H_{after}.

The noisy channel will now be analyzed mathematically in two different ways: (a) by the methods of communication theory, and (b) by the methods of mathematical psychology. The two methods will lead to identical results, but there is something to be learned by examining both methods. The "minimalist" reader may certainly skip over the next section and proceed directly to the section of psychological methods (The Noisy Channel II).

THE NOISY CHANNEL I:

The information about a transmitted ensemble of signals contained in the received ensemble of signals

Equation (4.1), which introduced us to the entropy concept, is a weighted average of the logarithms of the probabilities of the possible outcomes. We shall now go back a step – actually recede to a step

more elementary than Equation (4.1) – and examine the individual possible outcomes, rather than their average. However, we shall carry out this examination within the context of a noisy channel.

Following Middleton (1960), let us generalize the example of Figure 4.3 to include the presence of noise. Again we revert to the case of only two possible signals, but now we allow the possibility of mistakes [Figure (4.5), Box 4.1]. x_1 and x_2 will designate the transmitted signals, y_1 and y_2 the corresponding received signals. As before, let $p(x_1) = 0.8$ and $p(x_2) = 0.2$. Various non-zero conditional probabilities, $p(x_j|y_k)$, selected arbitrarily, are indicated on the diagram:

$$\begin{aligned} p(y_1|x_1) &= \quad 5/8 \\ p(y_2|x_1) &= \quad 3/8 \\ &\quad \text{sum to 1 .} \end{aligned}$$

Similarly

$$\begin{aligned} p(y_1|x_2) &= \quad 1/4 \\ p(y_2|x_2) &= \quad 3/4 \\ &\quad \text{sum to 1 .} \end{aligned}$$

$p(y_k)$, the probability of receiving y_k, is given by the equation

$$\begin{aligned} p(y_k) &= p(x_1, y_k) + p(x_2, y_k) \\ &= p(y_k|x_1) \cdot p(x_1) + p(y_k|x_2) \cdot p(x_2) . \end{aligned} \tag{4.13}$$

Thus

$$\begin{aligned} p(y_1) &= (5/8)(0.8) + (1/4)(0.2) = 0.55 \\ p(y_2) &= (3/8)(0.8) + (3/4)(0.2) = 0.45 . \end{aligned}$$

The $p(y_k)$ are entered in Figure 4.5.

We now make the following definitions:

$$\mathscr{H}(x_j) = \log p(x_j) = \text{initial or } \textit{a priori} \text{ uncertainty of the receiver about occurrence of } x_j . \tag{4.14}$$

$$\mathscr{H}(x_j|y_k) = -\log p(x_j|y_k) = \text{final or } \textit{a posteriori} \text{ uncertainty of the receiver about occurrence of } x_j \text{ after } y_k \text{ has been received.} \tag{4.15}$$

For example, a "zero" was received (y_k), but was a "one" really transmitted (x_j)? Then introducing the idea expressed by Equation (2.4) and (4.12) that

$$\text{Information} = H_{\text{before}} - H_{\text{after}} ,$$

we have

$$\mathscr{I}(x_j|y_k) = \mathscr{H}(x_j) - \mathscr{H}(x_j|y_k) \tag{4.16}$$

$$= \log[p(x_j|y_k)/p(x_j)] \tag{4.17}$$

from Equations (4.14) and (4.15). $\mathscr{I}_m$ is known as the *mutual information* of x_j and y_k.

The purpose of the example of Figure 4.5 is to calculate $\mathscr{I}_m(x_j|y_k)$ for all pairs (j, k); therefore, we shall need values of $p(x_j|y_k)$. We can calculate this conditional probability using the values for $p(x_j)$, $p(y_k)$ and $p(y_k|x_j)$ that have been entered in the diagram. Equating the right-hand sides of Equations (4.10) and (4.11),

$$p(x_j|y_k) = p(y_k|x_j)p(x_j)/p(y_k) . \tag{4.18}$$

Box 4.1 Signals transmitted with probability $p(X)$ and received with probability $p(y)$

The values for $p(y)$ were obtained in the following manner.

$$p(y_k) = \sum_{j=1}^{2} p(x_j, y_k), \text{ therefore,}$$

$$p(y_1) = p(x_1, y_1) + p(x_2, y_1)$$

$$= p(y_1|x_1) \cdot p(x_1) + p(y_1|x_2) \cdot p(x_2) ,$$

by Equation (4.13),

$$= (5/8 \times 0.8) + (1/4 \times 0.2) = 0.55$$

$$p(y_2) = p(x_1, y_2) + p(x_2, y_2)$$

$$= p(y_2|x_1) \cdot p(x_1) + p(y_2|x_2) \cdot p(x_2)$$

$$= (3/8 \times 0.8) + (3/4 \times 0.2) = 0.45 .$$

Having obtained the $p(y_k)$ we can now calculate the $p(x_j|y_k)$.

$$p(x_1|y_1) = p(y_1|x_1) \cdot p(x_1)/p(y_1) = (5/8) \times (0.8)/(0.55) = 0.90\dot{9}$$

$$p(x_2|y_1) = p(y_1|x_2) \cdot p(x_2)/p(y_1) = (1/4) \times (0.2)/(0.55) = 0.09\dot{0}\dot{9}$$

$$p(x_1|y_2) = p(y_2|x_1) \cdot p(x_1)/p(y_2) = (3/8) \times (0.8)/(0.45) = 0.66\dot{6}$$

$$p(x_2|y_2) = p(y_2|x_2) \cdot p(x_2)/p(y_2) = (3/4) \times (0.2)/(0.45) = 0.33\dot{3}$$

We can now calculate the mutual informations from Equations (4.16) and (4.17).

$$\mathscr{I}_m(x_j|y_k) = \mathscr{H}(x_j) - \mathscr{H}(x_j|y_k)$$

$$= -\log p(x_j) + \log p(x_j|y_k) = \log \frac{p(x_j|y_k)}{p(x_j)}$$

$$\mathscr{I}_m(x_1|y_1) = \log \frac{p(x_1|y_1)}{p(x_1)} = \log(0.909/0.8) = 0.128 \text{ n.u.}$$

$$\mathscr{I}_m(x_2|y_1) = \log \frac{p(x_2|y_1)}{p(x_2)} = \log(0.0909/0.2) = -0.788 \text{ n.u.}$$

$$\mathscr{I}_m(x_1|y_2) = \log \frac{p(x_1|y_2)}{p(x_1)} = \log(0.666/0.8) = -0.182 \text{ n.u.}$$

$$\mathscr{I}_m(x_2|y_2) = \log \frac{p(x_2|y_2)}{p(x_2)} = \log(0.333/0.2) = 0.511 \text{ n.u.}$$

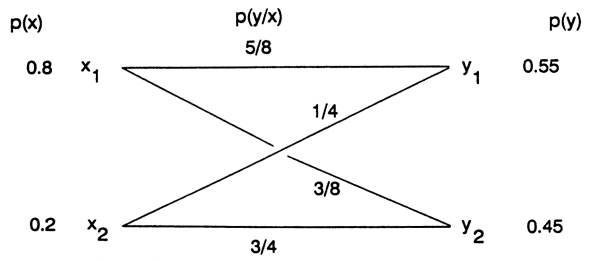

Figure 4.5 Information transmission in a noisy channel (cf. Figure 4.3 for a noiseless channel). The probabilities 5/8, 1/4, 3/8, 3/4 are conditional probabilities, $p(Y|X)$. The values for $p(y)$ are calculated from $p(X)$ and $p(Y|X)$ in the text.

The details of the calculations are given in Box 4.1, and the results are summarized below:

$$\mathscr{I}_m(x_1|y_1) = 0.128 \text{ natural units (n.u.)}$$
$$\mathscr{I}_m(x_2|y_1) = -0.788 \text{ n.u.}$$
$$\mathscr{I}_m(x_1|y_2) = -0.182 \text{ n.u.}$$
$$\mathscr{I}_m(x_2|y_2) = 0.511 \text{ n.u.}$$

We notice a strange phenomenon: $\mathscr{I}_m(x_2|y_1)$ and $\mathscr{I}_m(x_1|y_2)$ have negative values. The receiver has obtained negative information; his uncertainty about the transmitted signal is *greater* having received a signal than it was before the signal was transmitted. This result occurs when the conditional probability for x, for example $p(x_2|y_1)$, is less than the original probability for x, $p(x_2)$.

The example of Figure 4.5 can be expanded to include $n > 2$ possible signals.

The mutual information, $\mathscr{I}_m(x_j|y_k)$, is interesting heuristically, but our primary concern is the *average* information gain by the receiver. In the case of the noiseless channel we dealt with this matter by taking the *expectation*, E, of the logs of the transmission probabilities (see for example, Freund and Walpole [1980] or any standard text on probability),

$$H = -E[\log p_i] = -\sum_{i=1}^{n} p_i \log p_i \ . \tag{4.1}$$

We shall approach the noisy channel in the same way, *mutatis mutandis*.

Let us define

$$H(X) = E[\mathscr{H}(X)] \ = -\sum_{j=1}^{n} p(x_j) \log p(x_j) \ . \tag{4.19}$$

$H(X)$, the *source entropy* (cf. Equation (4.9) above), is the average *a priori* uncertainty about occurrence of x (before any x_j occurs). Recalling the result of Equation (4.8), we see that $H(X)$ will be maximum when all $p(x_j)$ are equal.

In a similar fashion, we can define the *conditional entropy*, $H(X|y)$, for the set of transmitted x_j given the received set y_k:

$$H(X|y) = E_{xy}[\mathscr{H}(x|y)] \ = -\sum_{j=1}^{n} \sum_{k=1}^{n} p(x_j, y_k) \log p(x_j|y_k) \ . \tag{4.20}$$

That is, the logarithms of the probabilities of x_j given y_k are averaged by means of a weighted sum of the joint probabilities of the occurrence of x_j and y_k. $H(X|y)$ is, then, the average *a posteriori* uncertainty about the set of x_j after the set y_k has been received. It represents the information lost in transmission and is sometimes known as the *equivocation*.

Again using the principle expressed by Equations (2.4), (4.12) and (4.16) we write

$$\boxed{\mathscr{I}(X|y) = H(X) - H(X|y) \ .} \tag{4.21}$$

$\mathscr{I}(X|y)$ is the average mutual information; or the average information about the ensemble of transmitted signals, X, contained in the ensemble of received signals, Y; or the average transmitted information. It will be referred to simply as the *transmitted information*. Expressing Equation (4.21) in words,

transmitted information = source entropy – equivocation

= source entropy – information loss due to errors in transmission.

Equation (4.21) is an explicit form of Equation (4.12). Written in full,

$$\boxed{\mathscr{I}(X|y) = -\sum_{j=1}^{n} p(x_j) \log p(x_j) + \sum_{j=1}^{n} \sum_{k=1}^{n} p(x_j, y_k) \log p(x_j|y_k) \ .} \tag{4.21a}$$

An alternative formulation of this equation is (see Appendix)

$$\mathscr{I}(X|y) = -\sum_{j=1}^{n} p(x_j)\log p(x_j) + \sum_{k=1}^{n} p(y_k)\sum_{j=1}^{n} p(x_j|y_k)\log p(x_j|y_k)\,. \tag{4.21b}$$

It can be shown that $\mathscr{I}(X|y)$ is always equal to or greater than zero.[1] When the channel is noiseless, $p(x_j, y_k) = 0$ for all $j \neq k$, so that from Equation (4.20), $H(X|y) = 0$, and Equation (4.21) becomes effectively identical to (4.1).

In the same way we define

$$H(X, y) = -\sum_{j=1}^{n}\sum_{k=1}^{n} p(x_j, y_k)\log p(x_j, y_k) \tag{4.22}$$

where $H(X, y)$ is the *joint entropy* of the ensembles. Analogously we can define the *receiver entropy*, $H(y)$, and the information $\mathscr{I}(Y|X)$. Various interesting relationships among these variables emerge, for which the reader is referred to the standard textbooks on communication theory.

The derivation of Equation (4.21) was our primary aim in this section. We shall now reorient ourselves and examine the noisy channel from the point of view of the psychologist of the 1950's.

THE NOISY CHANNEL II:

The information required for a categorical judgment. The "confusion" matrix

It was but three years after the appearance of Shannon's seminal work that Garner and Hake (1951) published their well-known paper entitled "The Amount of Information in Absolute Judgments." We shall derive Equation (4.21) again, but now within the context of a Garner-Hake experiment involving human judgments, rather than by analysis of the transmission of signals through a channel. Actually, there is no salient difference between these two paradigms, but it is instructive to look at the same problem in a different way.

The idea of an experiment involving judged categories can be illustrated with the following simple example. Suppose that there are three rods whose lengths are 10 cm, 20 cm and 30 cm. A subject is shown the 10 cm rod and told that it has a length of 10 cm; he or she is then shown the 20 cm rod and told that it has a length of 20 cm, etc. Thereafter, the experimenter draws the rods from the table and shows them to the subject, who tries to identify the rod as 10-, 20- or 30-cm by visual examination. The subject is not permitted to measure the rods. However, the chances are good that he or she will never make a mistake; the 10-cm rod will always be identified as 10 cm long, etc. Let us call the actual rod presented to the subject *the stimulus* (the 10-cm rod will be stimulus 1, the 20-cm rod, stimulus 2, and the 30-cm rod, stimulus 3), and let us call the subject's identification or reply, *the response* (the identification *this rod is rod* 1 will be *response* 1, etc.) Suppose the experiment continues until 100 stimulus-responses have been made. We can represent the stimuli by x_j, $j = 1, 2, 3$, and the responses by y_k, $k = 1, 2, 3$. That is, x_2 will represent a trial where the subject is presented with the 20-cm rod for identification, and y_3 will represent the act of the subject in identifying the 30-cm rod. The results of such an experiment will probably look much like those shown in Table 4.1. In this example the 10-cm rod stimulus was given 33 times ($x_1 = 33$) and correctly identified 33 times ($y_1 = 33$). It was never incorrectly identified as the 20-cm rod or the 30-cm rod. The 20-cm rod was given 34 times, etc. The columns and rows are each summated and, of course, the sum of sums for rows and the sum of sums for columns are each equal to 100. In the *stimulus-response matrix* depicted by the Table 4.1 only the diagonal elements are non-zero.

Suppose that the three rods are now cut to 20 cm, 22 cm and 24 cm, and the experiment is repeated. The subject is now going to make mistakes in identification. A possible stimulus-response matrix is shown in Table 4.2. In this hypothetical experiment the stimulus x_1, the 20-cm rod, was presented 33 times in all, was correctly identified only 21 times, was identified as y_2 on 7 occasions and as y_3 on 5 occasions. The stimulus-response matrix is no longer diagonal. The non-zero off-diagonal elements represent mistakes in identification (cf. errors in signal transmission). It would seem clear that a

Table 4.1 Three rods of lengths 10, 20 and 30 cm

	Response Categories, y_k			
Stimulus categories, x_j	y_1 = 10 cm	y_2 = 20 cm	y_3 = 30 cm	Totals
x_1 = 10 cm	33	0	0	33
x_2 = 20 cm	0	34	0	34
x_3 = 30 cm	0	0	33	33
Totals	33	34	33	100

stimulus-response matrix (sometimes called a "confusion" matrix) is, in principle, the same as a transmission-receipt matrix for a standard communication channel such as a telephone; in place of "stimulus" read "signal transmitted" and in place of "response" read "signal received."

There are three restrictions to the type of category experiment with which we shall be dealing. First, we are concerned only with those experiments dealing with stimuli of the "intensity" type, such as the intensity of light or of sound or the concentration of a solution or the magnitude of a force. That is, we shall *not* be concerned with stimuli such as the length of rods, although they served as a simple introduction to the confusion matrix. Second, we are interested primarily in the set of stimuli that span the totality of the physiological range, from threshold to maximum non-painful stimulus; for example auditory stimuli will extend from threshold to about 10^{10} or 10^{11} times threshold. The upper limit to the range of stimuli is often hard to define. Third, we are concerned primarily with the results obtained from "trained" subjects; that is, with subjects who have had as much time as desired to practice and learn which stimulus corresponds to which category. We speak more about the design of these experiments after we discuss the methods of analysis.

Let us now generalize the discussion of the stimulus-response matrix. Except for some minor changes in nomenclature, we follow the method of Garner and Hake. Although the number of stimulus categories need not, in principle, be equal to the number of response categories, we take them to be equal, just for simplicity.

Let N be the total number of trials, or the number of times a stimulus was presented to a given subject in the course of a single experiment. Then N_{jk} is the number of times a stimulus in category j was given and identified to be response category k. That is, N_{35} is the number of times stimulus category 3 was given by the investigator but identified or judged (incorrectly) to be (response) category 5. The N_{jk} can be tabulated as in Table 4.3. The sum of all elements in the k^{th} column equals $N_{\cdot k}$. That is,

$$\sum_{j=1}^{n} N_{jk} = N_{\cdot k} \quad . \tag{4.23}$$

Table 4.2 Three rods of lengths 20, 22 and 24 cm

	Response Categories, y_k			
Stimulus categories, x_j	y_1 = 20 cm	y_2 = 22 cm	y_3 = 24 cm	Totals
x_1 = 20 cm	21	7	5	33
x_2 = 22 cm	8	24	2	34
x_3 = 24 cm	5	9	19	33
Totals	34	40	26	100

Similarly, the sum of all elements in the j^{th} row equals $N_j.$. That is,

$$\sum_{k=1}^{n} N_{jk} = N_j. \quad . \tag{4.24}$$

The total number of stimuli given is equal to the total number of responses made:

$$\sum_{k=1}^{n} N_{\cdot k} = \sum_{j=1}^{n} N_j. = N \quad . \tag{4.25}$$

We can define the joint probability $p(x_j, y_k)$ by

$$p(x_j, y_k) = N_{jk}/N \quad . \tag{4.26}$$

We can also define the following probabilities, *a posteriori*, using $N_{\cdot k}$, $N_j.$ and N :

$$p(x_j) = N_j./N \tag{4.27}$$

$$p(y_k) = N_{\cdot k}/N \, . \tag{4.28}$$

Similarly, we can define the conditional probabilities

$$p(x_j|y_k) = N_{jk}/N_{\cdot k} \quad , \tag{4.29}$$

which is the conditional probability of stimulus x_j given response y_k; and

$$p(y_k|x_j) = N_{jk}/N_j. \quad , \tag{4.30}$$

which is the conditional probability of response y_k given stimulus x_j. Each of the equations (4.26), (4.29) and (4.30) provides the elements for a new matrix, which are shown in Tables 4.4a - 4.4c. We observe that all of the elements in the matrices shown in Tables 4.3 and 4.4 can be evaluated from the data collected in an experiment on categorical judgments performed in the manner described above.

Recalling again the two defining equations for conditional probability,

$$p(x_j, y_k) = p(x_j|y_k) \cdot p(y_k) \tag{4.10}$$

$$p(x_j, y_k) = p(y_k|x_j) \cdot p(x_j) \, . \tag{4.11}$$

The above two equations are verified by the *a posteriori* Equations (4.26) to (4.30). For example, using Equations (4.28) and (4.29), we can evaluate the right-hand side of Equation (4.10):

$$p(x_j|y_k) \cdot p(y_k) = \frac{N_{jk}}{N_{\cdot k}} \cdot \frac{N_{\cdot k}}{N} = \frac{N_{jk}}{N} \, .$$

By Equation (4.26), the left-hand side of (4.10) is given by N_{jk}/N, as required.

Table 4.3 Generalized Stimulus-Response Matrix

	Response categories				
Stimulus categories	y_1	y_2	y_k	y_n	Total
x_1	N_{11}	N_{12}	N_{1k}	N_{1n}	$N_1.$
x_2	N_{21}	N_{22}	N_{2k}	N_{2n}	$N_2.$
x_j	N_{j1}	N_{j2}	N_{jk}	N_{jn}	$N_j.$
x_n	N_{n1}	N_{n2}	N_{nk}	N_{nn}	$N_n.$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot k}$	$N_{\cdot n}$	N

Table 4.4a Dividing Each Element of the Stimulus-Response Matrix in Table 4.3 by N [Equation (4.26)] Produces a Matrix of Joint Probabilities, $p(x_j, y_k)$.

	Response categories		
Stimulus categories	y_1	y_k	y_n
x_1	$p(x_1, y_1)$	$p(x_1, y_k)$	$p(x_1, y_n)$
x_2	$p(x_2, y_1)$	$p(x_2, y_k)$	$p(x_2, y_n)$
x_j	$p(x_j, y_1)$	$p(x_j, y_k)$	$p(x_j, y_n)$
x_n	$p(x_n, y_1)$	$p(x_n, y_k)$	$p(x_n, y_n)$

Table 4.4b Dividing Each Element of the Stimulus-Response Matrix in Table 4.3 by $N_{.k}$ [Equation (4.29)] Produces a Matrix of Conditional Probabilities, $p(x_j|y_k)$.

	Response categories		
Stimulus categories	y_1	y_k	y_n
x_1	$p(x_1\|y_1)$	$p(x_1\|y_k)$	$p(x_1\|y_n)$
x_2	$p(x_2\|y_1)$	$p(x_2\|y_k)$	$p(x_2\|y_n)$
x_j	$p(x_j\|y_1)$	$p(x_j\|y_k)$	$p(x_j\|y_n)$
x_n	$p(x_n\|y_1)$	$p(x_n\|y_k)$	$p(x_n\|y_n)$

Table 4.4c Dividing Each Element of the Stimulus-Response Matrix in Table 4.3 by $N_{j.}$ [Equation (4.30)] Produces a Matrix of Conditional Probabilities, $p(y_k|x_j)$.

	Response categories		
Stimulus categories	y_1	y_k	y_n
x_1	$p(y_1\|x_1)$	$p(y_k\|x_1)$	$p(y_n\|x_1)$
x_2	$p(y_1\|x_2)$	$p(y_k\|x_2)$	$p(y_n\|x_2)$
x_j	$p(y_1\|x_j)$	$p(y_k\|x_j)$	$p(y_n\|x_j)$
x_n	$p(y_1\|x_n)$	$p(y_k\|x_n)$	$p(y_n\|x_n)$

We require, now, one final definition. Let

$$P(x_j, y_k) = p(x_j) \cdot p(y_k) \; . \tag{4.31}$$

We can interpret the quantity $P(x_j, y_k)$ as the probability of occurrence of two independent events. These independent events occur with probabilities $p(x_j)$ and $p(y_k)$. For example, if a coin is tossed on two occasions, the outcomes of the two tosses are independent. The probability of heads on the first toss is $\frac{1}{2}$ and the probability of heads on the second toss is $\frac{1}{2}$. Therefore the probability of heads on both tosses equals $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. However, the outcome y_k is, in general, not independent of the outcome x_j. That is, response y_k is not, in general, independent of the applied stimulus, x_j. In fact, we believe that the outcome of stimulus events and response events are quite closely related. If x_j and y_k were independent or "uncoupled" for some observer, $P(x_j, y_k)$ would give the probability of concurrence of the two outcomes. He or she would, however, be the poorest possible observer, since his or her responses would be totally unrelated to the corresponding stimulus (see, however, the problem for the reader in Chapter 5, Psychology: Categorical Judgments).

Now, we have defined the joint entropy of this stimulus-response system by Equation (4.22)

$$H(X, y) = -\sum_{j=1}^{n} \sum_{k=1}^{n} p(x_j, y_k) \log p(x_j, y_k) \; . \tag{4.22}$$

Therefore, we can define the *maximum* joint entropy by

$$H(X, y)_{\max} = -\sum_{j=1}^{n}\sum_{k=1}^{n} P(x_j, y_k)\log P(x_j, y_k)\,. \tag{4.23}$$

The maximum entropy corresponds to the most "disordered" system, which, in turn, corresponds to the system in which input (stimulus) and output (response) are totally uncorrelated. In such a system joint entropy is equal to $H(X, y)_{\max}$. In such a system no information about the stimulus is transmitted to the subject since he or she does not associate a given response with any particular stimulus.

We can now present an alternative to Equations (4.12) and (4.21), which stated that the average information about the ensemble of transmitted signals or stimuli contained in the ensemble of received signals or responses is given by

$$\boxed{\mathscr{I}(X|y) = H(X) - H(X|y)} \tag{4.21}$$

$$= H_{\text{before}} - H_{\text{after}}\,.$$

We now introduce as an equivalent mathematical statement

$$\boxed{\mathscr{I}(X|y) = H(X, y)_{\max} - H(X, y)\,.} \tag{4.33}$$

The quantities on the right-hand side can be evaluated using Equation (4.22) and (4.32) which, in turn, can be evaluated from the measured results. Equation (4.33) is, perhaps, more easily understood intuitively than (4.21). When $H(X, y)$ takes on its maximum value (transmitted and received signals independent), the transmitted information, $\mathscr{I}(X|y)$, equals zero as required. When $H(X, y)$ takes on its minimum value, $\mathscr{I}(X|y)$ is maximum. But $\mathscr{I}(X|y)$ is maximum when no errors in identification are made; that is when the matrix elements $N_{jk} = 0$ for $j \neq k$. We can see from Table 4.1 that under these conditions, Equation (4.33) reduces to

$$\mathscr{I}(X|y) = H(X)\,. \tag{4.34}$$

With reference now to Equation (4.21), we see that $\mathscr{I}(X|y)$ is maximum when the equivocation, $H(X|y)$, is equal to zero, when we have

$$\mathscr{I}(X|y) = H(X)\,. \tag{4.34}$$

So we obtain the same asymptotic result from (4.21) and (4.33). Therefore, on first glance, Equation (4.21) and (4.33) seem to exhibit similar properties. The interested reader is referred to the Appendix for a detailed proof that these equations are actually identical.

Equation (4.21) may be written in words:

$$\text{Information transmitted} = \text{stimulus entropy} - \text{stimulus equivocation.} \tag{4.35}$$

We may now add a symmetrical equation,

$$\mathscr{I}(Y|X) = H(y) - H(Y|X) \tag{4.36}$$

or, expressed in words,

$$\text{Information transmitted} = \text{receiver entropy} - \text{receiver equivocation.} \tag{4.37}$$

SUMMARY

Amid the maelstrom of equations in this chapter, let us keep in mind that what we have done is, in the final analysis, very simple. We have demonstrated in two ways (using the paradigm of a transmission line and of an experiment on categorical judgments) that information transmitted for a

noisy channel, $\mathscr{I}(X|y)$, may be calculated from Equations (4.21) and (4.36):

$$\begin{aligned}\mathscr{I}(X|y) &= H(X) - X(X|y)\\ &= H(y) - H(Y|X) \text{ bits per signal or bits per stimulus.}\end{aligned}$$

Some of the probabilities used to calculate $\mathscr{I}(X|y)$ may be known *a priori*, such as the $p(x_j)$, the probability of transmission of a signal or of application of a stimulus. Other probabilities may only be known *a posteriori*, after an experiment has been conducted. We shall run through an example of the calculation of $\mathscr{I}(X|y)$ in the next chapter.

The above equations have been derived from two ostensibly distinct, but nonetheless equivalent, starting points. We began with

$$\mathscr{I}(X|y) = H_{\text{before}} - H_{\text{after}} , \tag{4.12}$$

and alternatively with

$$\mathscr{I}(X|y) = H(X, y)_{\max} - H(X, y) . \tag{4.33}$$

Both viewpoints led to the same conclusion.

We note, finally, that when the equivocation, $H(X|y)$, is equal to zero, we obtain

$$\mathscr{I}(X|y) = H(X) , \tag{4.34}$$

which is the information transmitted for a noiseless channel, as expressed by Equation (4.1).

THE BOTTOM LINE

In order to calculate the average mutual information, $\mathscr{I}(X|y)$, one may use either Equation (4.21a) or (4.21b). Either of these equations may be conveniently utilized in a computer program, such as the one given in Chapter 5.

APPENDIX: THE EQUIVALENCE OF EQUATIONS (4.21) AND (4.33)

To demonstrate that

$$\mathscr{I}(X|y) = H(X) - H(X|y) \tag{4.21}$$

and

$$\mathscr{I}(X|y) = H(X, y)_{\max} - H(X, y) \tag{4.33}$$

are identical.

Beginning with Equation (4.33), we expand the right-hand side using Equations (4.22) and (4.32):

$$\begin{aligned}\mathscr{I}(X|y) = &-\sum_{j=1}^{n}\sum_{k=1}^{n} P(x_j, y_k)\log P(x_j, y_k)\\ &+\sum_{j=1}^{n}\sum_{k=1}^{n} p(x_j, y_k)\log p(x_j, y_k) .\end{aligned} \tag{A4.1}$$

The first double summation on the right-hand side can be simplified by introducing Equation (4.31), the defining equation for $P(x_j, y_k)$:

$$\begin{aligned}-\sum_{j=1}^{n}\sum_{k=1}^{n} P(x_j, y_k)\log P(x_j, y_k) &= -\sum_{j=1}^{n}\sum_{k=1}^{n} p(x_j)p(y_k)\log[p(x_j)p(y_k)]\\ &= -\sum_{j=1}^{n}\sum_{k=1}^{n} p(x_j)p(y_k)\log p(x_j) - \sum_{j=1}^{n}\sum_{k=1}^{n} p(x_j)p(y_k)\log p(y_k)\\ &= -\sum_{j=1}^{n} p(x_j)\log p(x_j)\sum_{k=1}^{n} p(y_k) - \sum_{j=1}^{n} p(x_j)\sum_{k=1}^{n} p(y_k)\log p(y_k) .\end{aligned}$$

Since $\sum_{j=1}^{n} p(x_j) = 1 = \sum_{k=1}^{n} p(y_k)$,

$$
\begin{aligned}
&-\sum_{j=1}^{n}\sum_{k=1}^{n} P(x_j, y_k)\log P(x_j, y_k) \\
&= -\sum_{j=1}^{n} p(x_j)\log p(x_j) - \sum_{k=1}^{n} p(y_k)\log p(y_k) \\
&= H(X) + H(y) \qquad \text{(A4.2)}
\end{aligned}
$$

by Equation (4.19) and its analog.

Continuing with the second double summation on the right-hand side of Equation (A4.1),

$$
\begin{aligned}
&\sum_{j=1}^{n}\sum_{k=1}^{n} p(x_j, y_k)\log p(x_j, y_k) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} p(y_k)p(x_j|y_k)\log[p(y_k)p(x_j|y_k)]
\end{aligned}
$$

using Equation (4.11),

$$
\begin{aligned}
&= \sum_{j=1}^{n}\sum_{k=1}^{n} p(y_k)\log p(y_k)\cdot p(x_j|y_k) + \sum_{j=1}^{n}\sum_{k=1}^{n} p(y_k)p(x_j|y_k)\log p(x_j|y_k) \qquad \text{(A4.3)} \\
&= \sum_{k=1}^{n} p(y_k)\log p(y_k)\sum_{j=1}^{n} p(x_j|y_k) + \sum_{j=1}^{n}\sum_{k=1}^{n} p(y_k)p(x_j|y_k)\log p(x_j|y_k) \\
&= \sum_{k=1}^{n} p(y_k)\log p(y_k) + \sum_{j=1}^{n}\sum_{k=1}^{n} p(x_j, y_k)\log p(x_j|y_k) \qquad \text{(A4.4)}
\end{aligned}
$$

(since $\sum_{j=1}^{n} p(x_j|y_k) = 1$)

$$
= -H(y) - H(X|y)\,, \qquad \text{(A4.5)}
$$

where

$$
H(y) = -\sum_{k=1}^{n} p(y_k)\log p(y_k) = \text{receiver entropy} \qquad \text{(A4.6)}
$$

and

$$
H(X|y) = -\sum_{j=1}^{n}\sum_{k=1}^{n} p(x_j, y_k)\log p(x_j|y_k) = \text{source equivocation.} \qquad (4.20)
$$

Combining Equations (A4.1), (A4.2) and (A4.5), we have

$$
\begin{aligned}
\mathscr{I}(X|y) &= H(X) + H(y) - H(y) - H(X|y) \\
\mathscr{I}(X|y) &= H(X) - H(X|y)\,. \qquad \text{(A4.7)/(4.21)}
\end{aligned}
$$

We observe that Equation (A4.7) is identical to Equation (4.21). This equation gives the average mutual information, or the average information about the

$$
\text{ensemble of} \left| \begin{matrix} \text{transmitted signals} \\ \text{stimuli} \end{matrix} \right| X\,,
$$

contained in the

$$
\text{ensemble of} \left| \begin{matrix} \text{received signals} \\ \text{responses} \end{matrix} \right| Y\,,
$$

or the *average transmitted information*.

Therefore, we have converted the right-hand side of Equation (4.33) into the right-hand side of Equation (4.21), showing that the equations are identical.

If we evaluate the second double summation on the right-hand side of Equation (A4.1) using the conditional probability $p(y_k|x_j)$, we can obtain in the same way Equation (4.36). Finally we should note that if we write

$$p(x_j, y_k) = p(y_k)\,p(x_j|y_k)\,, \tag{4.11}$$

introduce this quantity into Equation (4.20) and reverse the order of summation, we have

$$H(X|y) = -\sum_{k=1}^{n} p(y_k) \sum_{j=1}^{n} p(x_j|y_k) \log p(x_j|y_k) \tag{A4.8}$$

which is an alternative formulation of $H(X|y)$. This equation was used to derive Equation (4.21b).

NOTES

1. There are very few “It can be shown that” ’s in this book. I, personally, regard that phrase with a degree of suspicion: I suspect the author really can’t prove it, and this is his way of getting off the hook. But what really elevates suspicion to the point of certainty is when the author writes ”It can *easily* be shown that.” How many hours I have whiled away just to prove things that one author or another had found so obvious that it was not worth demonstrating! Would it have taken me that long if it could be “*easily shown*”? Certainly not! I’m sure the author is bluffing. All of them are. It can’t be me.

REFERENCES

Freund, J.E. and Walpole, R.E. 1980. *Mathematical Statistics*. Prentice-Hall, Englewood Cliffs, N.J.

Garner, W.R. and Hake, H. W. 1951. The amount of information in absolute judgments. *Psychological Review* **58**, 446 - 459.

Middleton, D. 1963. *An Introduction to Statistical Communication Theory*. McGraw-Hill, New York.

Raisbeck, G. 1963. *Information Theory: An Introduction for Scientists and Engineers*. M.I.T. Press, Cambridge.

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