CHAPTER 7

THE INFORMATION OF EVENTS WITH CONTINUOUS OUTCOMES

PROBABILITY DENSITY

We have studied in some detail the entropy of events with N possible *discrete* outcomes, whose *a priori* probabilities are $p_1, p_2, \ldots, p_n$:

$$H = -\sum_{i=1}^{n} p_i \log p_i \; . \tag{2.1}$$

The entropy of a potential winner of a lottery, when a single selection will be made from among m tickets which have been sold, is governed by the *a priori* probabilities,

$$p_{\text{Lisa}} = \text{(Number of tickets purchased by Lisa)}/m \; ,$$

$$p_{\text{Harry}} = \text{(Number of tickets purchased by Harry)}/m \; , \ldots$$

The identifying feature of this type of problem is that there are a finite number of outcomes (e.g. Lisa wins, Harry wins, ...), and an associated discrete *probability function*, defined by the *pure numbers* p_{Lisa}, p_{Harry}, ...

We can study the intensities (densities) of low levels of illumination by means of such discrete, probability functions. For example, the probability that one photon (light particle) will arrive at a rod (a type of light receptor) in some fixed interval of time is $p(\mathbf{x} = 1)$. The probability that two will arrive is $p(\mathbf{x} = 2)$, etc. The boldface character **x** signifies a *random variable* that may take on the values 1, 2, ... Equation (2.1) might be written

$$H = -\sum_{i} p(\mathbf{x} = x_i) \log p(\mathbf{x} = x_i) \; . \tag{7.1}$$

The probability function governing photon arrival is the Poisson function or distribution,

$$p(\mathbf{x} = z) = \frac{e^{-\lambda}\lambda^{z}}{z!} \; , \qquad z = 0, 1, 2, \ldots \tag{7.2}$$

which provides the probability that exactly z photons will arrive in a given interval of time. The parameter, λ, is the mean or average number of photons.

However, when the level of illumination becomes greater, photons are continually bombarding the retina of the eye, and intensity changes become, effectively, continuous. It is no longer possible to measure the effects of individual photons; we can no longer speak meaningfully of the probability that exactly $x_i = z$ photons will arrive. Rather we introduce a different kind of function called a *probability density function*, $p(x)$, such that $p(x)\Delta x$ is the probability that the intensity of light (measured on a continuous scale) lies between the values x and $x + \Delta x$, for small Δx. Statisticians do not seem overly fond of such definitions and prefer something like the following:

The probability that a random variable, **x**, will take on a value between a and b is given by

$$\int_a^b p(x)\, dx \; . \tag{7.3}$$

The density $p(x)$ will have the dimensions $[x^{-1}]$, so that $p(x)dx$ is a dimensionless probability. That is, $p(x)dx$ can be compared with the quantity $p(\mathbf{x} = x)$:

$$p(\mathbf{x} = x) \leftrightarrow p(x)\,\Delta x\,. \tag{7.4}$$

We can compare the normalization requirements for the two types of function. For the probability function,

$$\sum_i p(\mathbf{x} = x_i) = 1 \tag{7.5}$$

and for the probability density function $p(x) \geq 0$,

$$\int_{-\infty}^{\infty} p(x)\,dx = 1\,. \tag{7.6}$$

THE DIFFERENTIAL ENTROPY OF A PROBABILITY DENSITY FUNCTION

Well – if we can calculate the entropy of a (discrete) probability function using Equation (2.1), can we not simply substitute into Equation (7.1) replacing the $p(\mathbf{x} = x)$ by $p(x)\Delta x$? Let us take this approach and see.

Substituting relation (7.4) into Equation (7.1),

$$\begin{aligned} H &= -\sum_i p(x_i)\,\Delta x \log(p(x_i)\,\Delta x) \\ &= -\sum_i p(x_i)\log p(x_i)\,\Delta x - \left(\sum_i p(x_i)\,\Delta x\right)\log\Delta x\,. \end{aligned} \tag{7.7}$$

Then, since $\sum_i p(x_i)\Delta x = 1$, in the limit as $\Delta x \to 0$

$$H = -\int_{-\infty}^{\infty} p(x)\log p(x)\,dx - \lim_{\Delta x \to 0}\log\Delta x\,. \tag{7.8}$$

However, the second term on the right-hand side approaches negative infinity as $\Delta x \to 0$. That is, for any probability density function, $p(x)$, H will be infinite.

Equation (7.8) seems to be telling us that as the number of rectangles into which a continuous probability density function can be divided is increased progressively (refer Figure 7.1), the measure of "uncertainty" becomes larger without limit. Why does this phenomenon arise?

In order to understand the problem, it is useful to consider an example. Suppose that the probability density function $p(x)$ represents the height of adult males in some population. Let x be measured in centimeters. Then $p(x)\,\Delta x = p(175)(1)$ equals the probability that the height of a man selected at random from the population will be found to be between 175 and 175 + 1 cm. We could fragment the continuous probability density function into a number of narrow rectangles (say 250) with $\Delta x = 1$ cm, and calculate H from Equation (7.1):

$$H_1 = -\sum_{i=1}^{250} p(\mathbf{x} = x_i)\log p(\mathbf{x} = x_i)\,. \tag{7.9}$$

Two hundred and fifty centimeters is just a convenient upper limit for height. Then H_1 equals the value of the integral in Equation (7.8) minus log Δx, which is just equal to the value of the integral since $\Delta x = 1$. We could, however, carry the process further and divide the continuous probability density function into rectangles with width $\Delta x = 1$ mm $= 0.1$ cm., and evaluate H as

$$H_2 = -\sum_{i=0}^{2500} p(\mathbf{x} = x_i)\log p(\mathbf{x} = x_i)\,. \tag{7.10}$$

As before, we equate H_2 to the value of the integral in Equation (7.8) minus log Δx, and we find that H_2 is equal to the value of the integral minus log 0.1 , which is a greater entropy than before! We

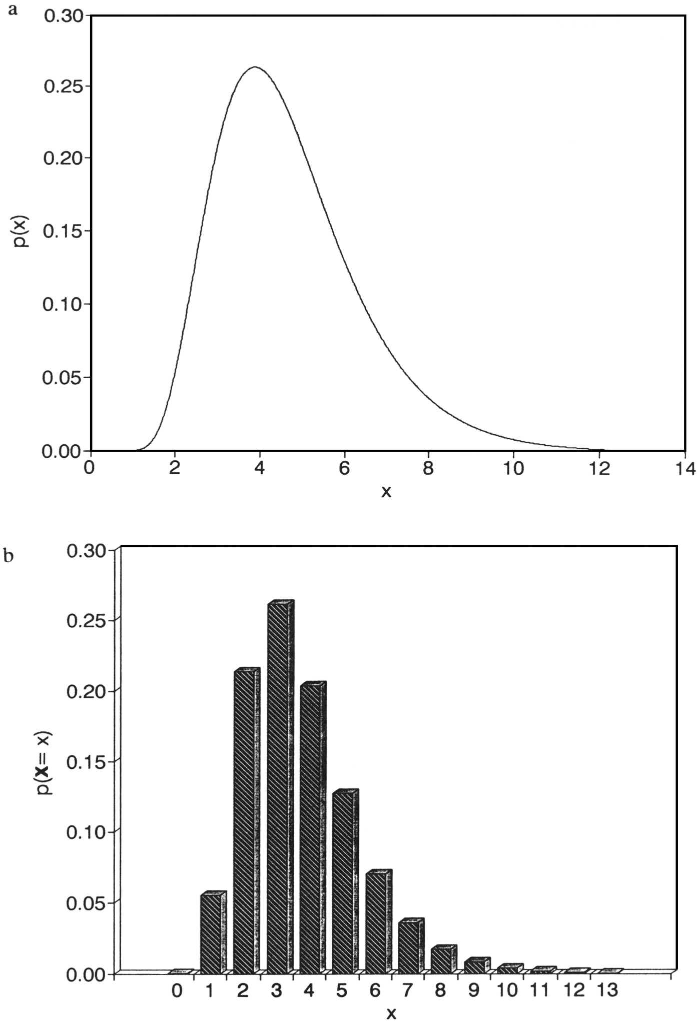

Figure 7.1 a & b The smooth curve in Fig 7.1a is the graph of the probability density function $p(x)$. The area under the curve is normalized to 1. One cannot, using the density function alone, obtain a proper measure of the information transmitted by making a measurement of the quantity, x. The probability function in Fig 7.1b was obtained by fragmenting the smooth density function into rectangles of unit width. The probability function is also normalized to 1: $\sum_{x=0}^{13} p(\mathbf{x} = x) = 1$. One can, indeed, obtain a measure of the information transmitted by making a measurement of the quantity, x, and using equation (7.1). However, the amount of information calculated will depend on the number of rectangles into which the smooth density function is divided.

appear to be more uncertain about the height of the next male sample who will be drawn from the population. That is,

$$H_2 > H_1 \,. \tag{7.12}$$

So the process would, apparently, continue. If $\Delta x = 0.1$ mm, H would increase again, and eventually reach infinity. The infinity arises because of the infinite number of choices we should have as $\Delta x \to 0$. As such, we receive an infinite quantity of information when the entropy was reduced by making a measurement of a man's height. It is essentially a question of resolution. We assumed first that a height of 176 cm is distinguishable from 175 cm; then we assumed that 175.1 cm is distinguishable from 175 cm; then we assumed that 175.01 cm is distinguishable from 175 cm, etc. If no limit is placed on the resolution of heights possible, then infinite information will be transmitted by any actual measurement of height. This argument can, perhaps, be made more concrete by use of a simple computer program.

In the following program, written in BASIC, a normal probability distribution with zero mean and unit variance has been taken. That is, $p(x) = N(x; 0, 1)$. This continuous probability density is first divided into rectangles of width 1, then of width 0.5, 0.25, ... Entropy is measured from the corresponding (discrete) histograms as in Equations (7.9) and (7.10), and is then calculated theoretically using Equation (7.8). We shall see in Chapter 8 that the integral in Equation (7.8) is equal to $\frac{1}{2}\ln(2\pi) + \frac{1}{2}$. It can be seen that the measured value of entropy agrees well with the theoretical value, and that fragmentation into rectangles still preserves the normalization of the density (area = 1). The value of entropy increases, apparently without limit, as the width of the rectangles decreases.

```
10 ’ PROGRAM IN BASIC TO ILLUSTRATE THAT DIVIDING A CONTINUOUS PROBABILITY
    DENSITY FUNCTION INTO A PROGRESSIVELY GREATER NUMBER OF RECTANGLES RESULTS IN
    A DISCRETE ENTROPY (H = -SUM P LOG P) THAT INCREASES WITHOUT LIMIT.
20 PRINT “ ENTROPY”, “ ENTROPY”, “ WIDTH OF”, “ AREA”
30 PRINT “(MEASURED)”, “(THEORETICAL)”, “ RECTANGLE”
40 ’
50 DEL = 1:’ WIDTH OF RECTANGLES BEGINS AT 1.
60 ’
70 ’ SELECT A GAUSSIAN FUNCTION WITH UNIT VARIANCE
80 DEF FNA(X) = (1 / SQR(2 * 3.1416)) * EXP(-X * X / 2)
90 ’
100 ’ CALCULATE THE AREA UNDER THE GAUSSIAN FUNCTION AND THE VALUE OF ENTROPY, H.
110 ’
115 WHILE 1
120 H = 0: AREA = 0
130 FOR X = -10 TO 10 STEP DEL
140 AREA = AREA + FNA(X) * DEL
150 H = H - DEL * FNA(X) * LOG(DEL * FNA(X))
160 NEXT X
170’
180 ’ THEORETICALLY, H EXCEEDS THE VALUE OF THE DIFFERENTIAL ENTROPY,
    (0.5 * LOG(2 * Pi) + 0.5), BY THE NEGATIVE LOG OF THE WIDTH OF THE RECTANGLE.
190 HTHEOR = 0.5 * LOG(2 * 3.1416) + 0.5 - LOG(DEL)
200 PRINT H, HTHEOR, DEL, AREA
210’
220 ’DECREASE THE WIDTH OF THE RECTANGLES BY A FACTOR OF 2 AND REPEAT...
230 DEL = DEL / 2
240 WEND

RUN

 ENTROPY      ENTROPY         WIDTH OF     AREA
 (MEASURED)   (THEORETICAL)   RECTANGLE
 1.418938     1.41894         1            0.9999989
 2.112085     2.112087        0.5          0.9999988
 2.805231     2.805234        0.25         0.9999988
 3.498378     3.498381        0.125        0.9999989
 4.191524     4.191528        0.0625       0.9999988
 4.88467      4.884676        0.03125      0.9999989
 5.577818     5.577823        0.015625     0.9999982
 6.270962     6.27097         0.0078125    0.9999981
 ..           .               .            .
 ..           .               .            .
```

From a practical point of view, the above argument of decreasing the value of Δx without limit is ludicrous. There is a limit to the precision with which the height of a person can be measured. Small changes in posture, fluctuations in intervertebral spacing, etc. produce such a limit. Probably the precision of the measurement is not as small as 0.1 mm. Therefore, there is little point to measuring H with Δx < 0.1 mm, and, therefore, we do not obtain an infinite quantity of information from a measurement of height.

There is nothing unique about our example of measuring heights. Every measurement of a quantity that is continuously distributed is limited in precision either by the nature of the measuring apparatus, or by the properties of the measured quantity itself. Therefore, neither H nor the amount of information received from the measurement will be infinite. Let us have another look at Equation (7.8).

The integral $-\int_{-\infty}^{\infty} p(x)\log p(x)\,dx$ that appears in Equation (7.8) is known as a *differential entropy* (McEliece, 1977). You may be thinking, as I did when I first encountered the differential entropy, that it is a natural extension of Equation (7.1) into the continuous domain, and would be a natural way of expressing the uncertainty associated with a probability density function. Life, however, is not this simple. The differential entropy, as we have seen, is only one of two expressions that evolve from the H-function for discrete random variables when $p(\mathbf{x} = x)$ is replaced by $p(x)\,dx$. Therefore, numerical values of the differential entropy are not, in themselves, a reflection of our earlier notions of uncertainty. The differential entropy may even turn out to have a negative value. Moreover, the differential entropy is not "coordinate-free"; its value will depend on the units in which x is measured. For example, if $p(x)$ is, again, the probability density function for heights of people, x may be measured in centimeters so that

$$H_{\text{centimeter}} = -\int_{-\infty}^{\infty} p(x)\log p(x)\,dx\ . \tag{7.12}$$

Suppose, however, that values of the random variable are measured in meters, and let $g(y)$ be the new probability density function for heights. Then, using the theorem for change of variable (refer, for example, Freund and Walpole, 1980),

$$g(y) = p(x)\left|\frac{dx}{dy}\right|\ . \tag{7.13}$$

But $x = 100\,y$ (that is, 1 meter = 100 centimeters), so that $\left|\frac{dx}{dy}\right| = 100$. Hence

$$g(y) = 100\,p(x)\ .$$

Then (leaving out the limits of integration)

$$\begin{aligned} H_{\text{meter}} &= -\int g(y)\log g(y)\,dy = -\int 100\,p(x)\log(100p(x))\,\frac{dx}{100} \\ &= -\int p(x)\log p(x)\,dx - \int p(x)\log(100)\,dx \end{aligned}$$

$$H_{\text{meter}} = H_{\text{centimeter}} - \log 100\ , \tag{7.14}$$

(Shannon and Weaver, 1964, page 91). That is, the differential entropy is neither a measure of uncertainty in keeping with our previous notions of uncertainty, nor is it even independent of the units of measurement of values of the random variable. This example is illustrated in Figure 7.2.

THE ABSOLUTE ENTROPY

We have been dealing with a probability density function, $p(x)$. Suppose we now consider some other probability density function, $q(x)$, defined on x. Treating $q(x)$ in exactly the same manner as $p(x)$, we obtain for its entropy using Equation (7.8),

$$H' = -\int q(x)\log q(x)\,dx - \lim_{\Delta x\to 0}\log\Delta x\ . \tag{7.15}$$

Subtracting H' from H,

$$H - H' = -\int p(x)\log p(x)\,dx - \left(-\int q(x)\log q(x)\,dx\right) \tag{7.16}$$

since the two Δx-terms cancel. $H - H'$, the difference between two differential entropies, is known as the *absolute entropy*. In contrast to the differential entropy, the absolute entropy *is* coordinate free. That is, $H_{\text{abs}} = H - H'$ will have the same value regardless of the units in which values of the random variable are measured. This may be illustrated from the centimeter-meter example of the previous

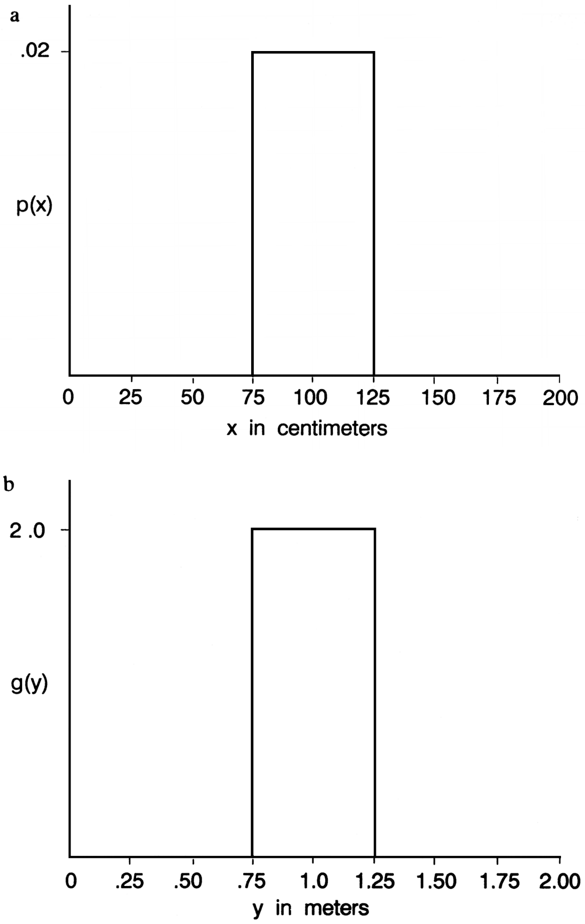

Figure 7.2 a & b We suppose that the probability density function for heights of adult males in some (strange) population is governed by the uniform distribution: $p(x) = 0.02,\ 75 \le x \le 125$, where x is measured in centimeters. We note that $p(x)$ is normalized so that $\int_{-\infty}^{\infty} p(x)dx = (0.02)(50) = 1.0$. Another density function governing the same random process is the uniform distribution $g(y)$, defined by $g(y) = 2.0,\ 0.75 \le y \le 1.25$, where y is measured in meters. Again, note that $g(y)$ is normalized so that $\int_{-\infty}^{\infty} g(y)dy = (2.0)(0.50) = 1.0$.
The differential entropy of $p(x)$ can be evaluated explicitly: $H_{\text{centimeter}} = -(0.02\ \log 0.02)(125-75) = -\log 0.02 = \log 50$. The differential entropy of $g(y)$ can also be evaluated explicitly: $H_{\text{meter}} = -(2.0\ \log 2.0)(1.25-0.75) = -\log 2.0 = \log 0.5$. That is, $H_{\text{meter}} = H_{\text{centimeter}} - \log 100$, as given by Equation (7.14). The above illustrates that differential entropy is not independent of the units in which the random variable is measured.

section. When measuring in units of centimeters,

$$H_{\substack{\text{abs}\\ \text{centimeter}}} = -\int p(x)\log p(x)\,dx + \int q(x)\log q(x)\,dx\,.$$

When measuring in units of meters,

$$\begin{aligned} H_{\substack{\text{abs}\\ \text{meter}}} &= \left[-\int p(x)\log p(x)\,dx - \log 100\right] - \left[\int q(x)\log q(x)\,dx - \log 100\right] \\ &= H_{\substack{\text{abs}\\ \text{centimeter}}}\,. \end{aligned}$$

For a more general proof refer, for example, to the text by S. Goldman (1953).

So the absolute entropy is coordinate free; but what, if anything, has it to do with the transmission of information? We recall that for a noisy channel it was necessary to represent the information transmitted per symbol as the difference between two H-functions:

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

That is, information was obtained as the difference between the uncertainty that existed before the measurement was made and the uncertainty that remained after the measurement had been made. The noise in the channel limited the accuracy with which it was possible to transmit a symbol. We now call upon this idea again in the attempt to adapt the absolute entropy for the measurement of information.

We recall from the example of measuring heights that the measurement was limited in its precision by various anatomical factors such as fluctuating intervertebral spacings. These factors limiting the precision of a measurement play the part of a noise that limits the amount of information transmitted. The probability density function for height may, therefore, be regarded as a density function for "pure" height compounded with "noise" (read, perhaps, as, "pure signal" with noise). In Equation (7.16) for absolute entropy, then, let us regard $p(x)$ as $p_{SN}(x)$ (that is, $p_{\text{signal-noise}}(x)$), and $-\int p_{SN}(x) \log p_{SN}(x)\, dx$ as the differential entropy of signal together with noise.

What role, then, is played by the differential entropy

$$-\int q(x) \log q(x)\, dx\, ?$$

A reasonable suggestion is that $q(x)$ be the density function for noise alone, which we might represent by $p_N(x)$. Then, by conjecture, the absolute entropy is equal to the information "content" of a measurement, by which we mean the information received by the measurer when the measurement is made. That is,

$$\mathscr{I} = H_{\text{abs}} = -\int p_{SN}(x) \log p_{SN}(x)\, dx + \int p_N(x) \log p_N(x)\, dx \, . \tag{7.17}$$

It is apparent that Equation (7.17) possesses at least one necessary asymptotic property. That is, as $p_{SN} \to p_N$ (pure signal vanishes), $\mathscr{I} = H_{\text{abs}} \to 0$, meaning that no information is received, which makes sense, because under these circumstances there is no discernible signal. Further investigation of Equation (7.17) as a measure of information will have to await a discussion of the means of evaluating p_{SN}.

We might speculate that what has been done by the subtraction process in Equation (7.17) is to convert a continuous probability density function, p_{SN} , which might be represented by Figure 7.1a (and from which we obtain, ostensibly, an infinite quantity of information) into a histogram of the type shown in Figure 7.1b, from which we can obtain a proper measure of information. The effect of the subtraction process in Equation (7.17) is to "discretize" the continuous density function. The width of the bars in the histogram are determined by the noise level, or the limitation to our measurement of values of x. If we represent Equation (7.17) as

$$\mathscr{I} = H_{SN} - H_N \tag{7.18}$$

we might compare it to the explicit form of the information equation for a discrete, noisy channel:

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

Although the comparison is not perfect, the equivocation, $H(X|Y)$, seems to play the part of the differential entropy of noise.

CONVOLUTION

It was stated above that $p_{SN}(x)$ is the probability density function for pure signal compounded with noise, without suggesting a procedure for obtaining it. The method used is one of convolution. Suppose that $p_S(x)$ and $p_N(x)$ are density functions for signal and noise respectively. We recall that $p_{SN}(x)\Delta x$ is the probability of finding a value of signal-plus-noise whose total lies between the values of x and $x + \Delta x$. Thus, for example, $p_{SN}(10)\Delta x$ is the probability that $x_S + x_N = 10$ very nearly. There are, however, infinitely many ways by which $x_S + x_N = 10$:

$$\begin{array}{lll} x_S = 14, & x_N = -4, & x_S + x_N = 10 \\ x_S = 13, & x_N = -3, & x_S + x_N = 10 \\ \ldots. & \ldots & \ldots \\ x_S = 10, & x_N = 0, & x_S + x_N = 10 \\ \ldots. & \ldots & \ldots \\ x_S = -5, & x_N = 15, & x_S + x_N = 10 \\ \ldots. & \ldots & \ldots \end{array}$$

Thus, if we let $x = x_S + x_N$ and $x' = x_N$, then $x_S = x - x'$. Therefore,

$$\overset{\substack{x_S + x_N \\ \downarrow}}{p_{SN}(x)\Delta x} = \sum_{x'=-\infty}^{\infty} p_S \overset{\substack{x_S \\ \downarrow}}{\overbrace{(x - x')}} \Delta x \cdot \overset{\substack{x_N \\ \downarrow}}{p_N(x')\Delta x'} \ . \tag{7.19}$$

The probabilities $p_S(x - x')\Delta x$ and $p_N(x')\Delta x'$ are multiplied to give the probability of concurrence of the two independent events, and the products are added to give the total probability of these mutually exclusive events. Taking Equation (7.19) to the limit as $\Delta x' \rightarrow 0$,

$$p_{SN}(x) = \int_{-\infty}^{\infty} p_S(x - x')\, p_N(x')\, dx' \ . \tag{7.20}$$

Equations (7.19) and (7.20) could equally well have been written with the arguments of the two functions interchanged. The integral in Equation (7.20) is known as a *convolution integral*, and it its found in many branches of mathematics.[1]

Using Equation (7.20), we can, in principle, obtain p_{SN} explicitly, by convolving the two functions p_S and p_N. The integration can often be carried out analytically.

We can now, in principle, evaluate $\mathscr{I}$, the information contained in a measurement made of a variable that is distributed continuously, from knowledge of the two probability density functions, p_S and p_N:

(a) Using Equation (7.20) we obtain $p_{SN}(x)$.
(b) Using Equation (7.17) / (7.18) we obtain $\mathscr{I}$.

APPLICATIONS IN PERCEPTUAL STUDIES

There are very, very few investigators who have found it desirable to use the information from "continuously distributed" probability density functions (sometimes referred to as "information from analog channels") in studies of perception. Such analog-information will, however, be required almost exclusively in the sensory studies that follow in this book. It is well to remember that while differential entropy is not a proper measure of information, any difference between two differential entropies evaluated for the same random variable, x, *is* indeed a correct measure of information. The differential entropy that is subtracted need not always represent the obfuscation of a "noise." For example, consider two absolute entropies

$$H_{\text{abs1}} = H_{\text{diff1}} - H_{\text{noise}} \tag{7.21}$$

$$H_{\text{abs2}} = H_{\text{diff2}} - H_{\text{noise}} \ . \tag{7.22}$$

Then

$$H_{\text{abs1}} - H_{\text{abs2}} = H_{\text{abs}} = H_{\text{diff1}} - H_{\text{diff2}} \,. \tag{7.23}$$

H_{abs} is found as the difference between two differential entropies, neither of which represents noise. We shall have occasion to obtain an absolute entropy as the difference between two differential entropies, the second of which will be labeled “reference”:

$$H_{\text{abs}} = H_{\text{diff}} - H_{\text{reference}} \,. \tag{7.24}$$

$H_{\text{reference}}$ may or may not be a noise in the usual sense, but it will, nonetheless, limit the precision with which a sensory measurement can be made. (Please refer also to Chapter 16.)

H_{abs} becomes a little more tangible when it is evaluated for specific probability density functions. The object of the next chapter will be to evaluate H_{abs} when the differential entropies from which it is obtained issue from the normal or Gaussian distribution.

NOTES

1. “Convolution” (Latin, *con*: together + *volvere*: to roll). Perhaps better is “folded back.” As the dummy variable x' increases, the argument of $p_{SN}(x')$ increases, while that of $p_S(x - x')$ decreases.

REFERENCES

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

Goldman, S. 1953. *Information Theory*. Prentice-Hall, Englewood Cliffs, N.J.

McEliece, R.J. 1977. *The Theory of Information and Coding: A Mathematical Framework for Communication*. Addison-Wesley, Reading, Mass.

Shannon, C.E. and Weaver, W. 1964. *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana.