c14g, 6/15/1994

CHAPTER 14

SIMPLE APPLICATIONS OF DECISION THEORY

We now examine in detail two of the simplest applications of the general decision theory just formulated, and compare the first with the older Neyman–Pearson procedure. The problem of detection of signals in noise is really the same as Laplace's old problem of detecting the presence of unknown systematic influences in celestial mechanics, and Shewhart's (1931) more recent problem of detecting a systematic drift in machine characteristics, in industrial quality control. Statisticians would call the procedure a "significance test." It is unfortunate that the basic identity of all these problems was not more widely recognized, because it forced workers in several different fields to rediscover the same things, with varying degrees of success, over and over again.

As is clear by now, all we really have to do to solve this problem is to take the principles of inference developed in Chapters 2 and 4; and supplement them with the loss function criterion for converting final probabilities into decisions (and, if needed, the maximum entropy principle for assigning priors). However, the literature of this field has been created largely from the standpoint of the original decision theory before this was realized. The existing literature therefore uses a different sort of vocabulary and set of concepts than we have been using up to now. Since it exists, we have no choice but to learn these terms and viewpoints if we want to read the literature of the field. This material appeared in the papers of Middleton and van Meter (1955, 1956), and later in the monumental treatise of Middleton (1960), in an enormously expanded form where a beginner can get lost for months without ever finding the real underlying principles. So we need a very rapid, condensed review of the literature of the 1950's on these problems. To have a complete, self–contained summary, we repeat a little from previous chapters as a way of introducing this different language.

Definitions and Preliminaries

Notation:

$$\begin{aligned} p(A|B) &= \text{ Conditional probability of } A\text{, given } B \\ p(AB|CD) &= \text{ Joint conditional probability of } A \text{ and } B\text{, given } D \text{ and } C \ldots \text{etc.} \end{aligned}$$

For our purposes, everything follows from the product rule:

$$p(AB|C) = p(A|BC)\,p(B|C) = p(B|A)\,p(A|C). \tag{14–1}$$

If the propositions B and C are not mutually contradictory, this may be rearranged to give the rule of "learning by experience," Bayes' theorem:

$$p(A|BC) = p(A|C)\,\frac{p(B|AC)}{p(B|C)} = p(A|B)\,\frac{p(C|AB)}{p(C|B)}. \tag{14–2}$$

If there are several mutually exclusive and exhaustive propositions B_i, then by summing (14–1) over them, we obtain the chain rule

$$p(A|C) = \sum_i p(A|B_iC)\,p(B_i|C) \tag{14–3a}$$

or, in a simple skeleton notation,

$$p(A|C) = \sum_B p(A|BC)\, p(B|C). \tag{14–3b}$$

Now let

$X =$ prior knowledge, of any kind whatsoever

$S =$ signal

$N =$ noise

$V = V(S, N) =$ observed voltage

$D =$ decision about the nature of the signal

Thus we have

$p(S|X) =$ prior probability of the particular signal S

$p(N|X) = W(N) =$ prior probability of the particular sample of noise N.

We understand that the prior information X is always built into the right–hand side of all our probability symbols, whether or not we write it explicitly. Thus, in a linear system, $V = S + N$, and

$$p(V|S) \equiv p(V|SX) = W(V - S). \tag{14–4}$$

The reader may be disturbed by the absence of density functions, dS's, dN's, *etc.*, which might be expected in the case of continuous S, N. Note, however, that our equations are homogeneous in these quantities, so they cancel out anyway. We are trying only to convey the broad ideas, without bothering with fine details which would make the notation very intricate. Thus by $\sum_A$ we mean ordinary summation over some previously agreed set of possible values if A is discrete, integration with appropriate density functions if A is continuous.

A *decision rule* $p(D_i|V_j)$, or for brevity just $p(D|V)$, represents the process of drawing inferences about the signal from the observed voltage. If it is always made in a definite way, then $p(D|V)$ has only the value 0, 1 for any choice of D and V; however we may also have a "randomized" decision rule according to which $p(D|V)$ is a true probability distribution. Maintaining this more general view turns out to be a help in formulating the theory.

The essence of any decision rule, and in particular, any one which can be built into automatic equipment, is that the decision must be made on the basis of V alone; V is, by definition, the quantity which contains all the information actually used (in addition to the ever-present X) in arriving at the decision. Thus, if $Y \neq D$ is any other proposition, we have

$$p(D|V) = p(D|VY). \tag{14–5}$$

The fact that Y is to be ignored in the presence of V might appear a departure from our previous exhortations that the robot is always to take into account all the relevant information it has. However, if we consider that the property (14–5) is a part of the prior information X there is no difficulty. To put it differently, (14–5) expresses the prior knowledge that there is a direct logical relation by which D is determined by V alone. If this relation was a known law of physics, there would be nothing strange in (14–5). The only difference is that in the present case this relation does not express any law of Nature, but rather our own design of the apparatus. In either case, Y is ignored not because the robot has relaxed its rules, but because Y is irrelevant.

An equivalent statement is that the probability of reaching a decision D depends on any proposition Y only through the intermediate influence of Y on V:

$$p(D|Y) = \sum_V p(D|V)\, p(V|Y). \tag{14–6}$$

which is a kind of "Huygens principle" for logic. To see the analogy, think of Y as a light source which cannot be seen from D, but it illuminates various points V. Then the resulting light arriving at D is the sum of the Huygens wavelets $p(D|V)$ with amplitudes $p(V|Y)$. The almost exact mathematical analogy between conditional information flow and the flow of light according to the Huygens principle of optics will appear in detail when we consider the statistical mechanics of irreversible processes.

Sufficiency and Information

Equation (14–5) has interesting consequences; suppose we wish to judge the plausibility of some proposition Y, on the basis of knowledge of V and D. From the product rule (14–1),

$$p(DY|V) = p(Y|VD)\,p(D|V) = p(D|VY)\,p(Y|V)$$

and using (14–5), this reduces to

$$p(Y|VD) = p(Y|V). \tag{14–7}$$

Thus, if V is known, knowledge of D is redundant and cannot help us in estimating any other quantity. The reverse is not true, however; we could equally well use (14–5) in another way:

$$p(VY|D) = p(Y|VD)\,p(V|D) = p(Y|D)\,p(V|YD).$$

Combining this with (14–7), there results the

Theorem: Let D be a possible decision, given V. Then $p(V|D) \neq 0$, and $p(Y|V) = p(Y|D)$ if and only if $p(V|D) = p(V|YD)$. (14–8)

In words: knowledge of D is as good as knowledge of V for judgments about Y if and only if Y is irrelevant for judgments about V, given D. Stated differently: in the "environment" produced by knowledge of D, the probabilities of Y and V are independent, *i.e.,*

$$p(YV|D) = p(Y|D)\,p(V|D). \tag{14–9}$$

In this case, in the literature of this field D is said to be a *sufficient statistic* for judgments about Y. We shall want to see whether this is in agreement with our earlier definitions of sufficiency, made from a quite different point of view in Chapter 8.

Evidently, a decision rule which makes D a sufficient statistic for judgments about the signal S is superior to one without this property, in that it tells us more about the signal. However, such a rule does not necessarily exist. Equation (14–9) is a very restrictive condition, since it must be satisfied for all values of Y, V, and all D for which $p(D|V) \neq 0$.

As you might guess from this, the concept of sufficiency is closely related to that of information. The above definition of sufficiency could be stated equally well as: D is a sufficient statistic for judgments about Y if it contains all the information about Y which V contains. Since D is determined from V, if it is not a sufficient statistic, it seems intuitively that is necessarily contains *less* information about Y than does V. In this statement, the term "information" was used in a loose, intuitive sense; does it remain true if we adopt Shannon's measure of information?

Imagine that there are several mutually exclusive propositions Y_i, one of which must be true. For brevity we use, as above, the notation $\sum_Y f(Y) \equiv \sum_i f(Y_i)$. With a specific value of D given, the entropy which measures our information about the propositions Y_i is

$$H_D(Y) = -\sum_Y p(Y|D) \log p(Y|D)$$

and its average over all values of D is

$$\overline{H}_D(Y) = \sum_D p(D|X)\, H_D(Y). \tag{14–11}$$

If

$$\overline{H}_C(Y) < \overline{H}_D(Y) \tag{14–12}$$

we say that C contains, on the average, more information about Y than does D. Note, however, that it may be otherwise for specific values of C and D.

Acquisition of new information can never increase $\overline{H}$; let $\{Z_i\}$ be, for the moment, any set of propositions and form the expression

$$\overline{H}_V(Z) - \overline{H}_{DV}(Z) = \sum_{DVZ} p(DV|X)\, p(Z|DV) \log p(Z|DV) - \sum_{VZ} p(V|X)\, p(Z|V) \log p(Z|V)$$

$$= \sum_{DVZ} p(CV|X)\, p(Z|DV) \log \left[\frac{p(Z|DV)}{p(Z|V)}\right].$$

Using the fact that on the positive real line $\log x \geq (1 - x^{-1})$, with equality if and only if $x = 1$, this becomes

$$\overline{H}_V(Z) - \overline{H}_{DV}(Z) \geq \sum_{DVZ} p(DV|X)[p(Z|DV) - p(Z|V)] = 0 \tag{14–13}$$

Thus, $\overline{H}_{DV}(Z) \leq \overline{H}_V(Z)$, with equality if and only if Eq. (14–7) holds for all D, V, and Z for which $p(DV|X) \neq 0$.

But now, since (14–13) holds regardless of the meaning of D and V, we can conclude equally well that for all D, V, Z,

$$\overline{H}_D(Y) \geq \overline{H}_{DV}(Z) \leq \overline{H}_V(Z).$$

Choosing $Z = Y$, we have in consequence of (14–7) $H_V(Y) = H_{DV}(Y)$, so that

$$\overline{H}_V(Y) \leq \overline{H}_D(Y) \tag{14–14}$$

with equality if and only if Eq. (14–9) holds, *i.e.,* if and only if D is a sufficient statistic as just defined. Thus, if by "information" we mean minus the average entropy of Y over the prior distribution of D or V, zero information loss in going from V to D is equivalent to sufficiency of D.

Note that inequality (14–13) holds only for the averages $\overline{H}$, not for the H. Acquisition of a specific piece of information (that an event previously considered improbable had in fact occurred) may in some cases increase the entropy of Y. However, this is an improbable situation and on the average the entropy can only be lowered by additional information. This shows again that the term "information" is not a happy choice of words to describe entropy expressions. In spite of the entropy increases, the situation just described could hardly be called one of less *information* in the colloquial sense of that word; but rather one of less *certainty.*

Loss Functions and Criteria of Optimum Performance

In order to say that one decision rule is better than another, we need some specific criterion of what we want our detection system to accomplish. The criterion will vary with the application, and obviously no single decision rule can be best for all purposes. But our discussion in Chapter 13 will apply, almost unchanged, in this slightly different language. A very general type of criterion is obtained by assigning a *loss function* $L(D, S)$ which represents our judgment of how serious it is to make decision D when signal S is in fact present.

In case there are only two possible signals, $S_0 = 0$ (*i.e.,* no signal), and $S_1 > 0$, and consequently two possible decisions D_0, D_1 about the signal, there are two types of error, the false alarm $A = (D_1, S_0)$ and the false rest $R = (D_0, S_1)$. In some applications, one type of error might be much more serious than the other.

Suppose that a false rest is considered ten times as serious as is a false alarm, while a correct decision of either type represents no "loss." We could then take $L(D_0, S_0) = L(D_1, S_1) = 0$, $L(D_0, S_1) = 10$, $L(D_1, S_0) = 1$. Whatever the possible signals and the possible decisions form discrete sets, the loss function becomes a *loss matrix.* In the above example,

$$L_{ij} = \begin{pmatrix} 0 & 10 \\ 1 & 0 \end{pmatrix}$$

Instead of assigning arbitrary a certain loss value to each possible type of detection error, we may consider *information loss* by the assignment $L(D, S) = -\log p(S|D)$. This is somewhat more difficult to manipulate, because now $L(D, S)$ depends on the decision rule. A decision rule which minimizes information loss is one which makes the decision in some sense as close as possible to being a sufficient statistic for judgments about the signal. In exactly what sense seems never to have been clarified.

The *conditional loss* $L(S)$ is the average loss incurred when the specific signal S is present

$$L(S) = \sum_D L(D, S)\, p(D|S) \tag{14–15}$$

which may in turn be expressed in terms of the decision rule and the properties of the noise by using (14–6). The *average loss* is the expected value of this over all possible signals:

$$\langle L \rangle = \sum_S L(S)\, p(S|X). \tag{14–16}$$

Two different criteria of optimum performance now suggest themselves:

The Minimax Criterion. For a given decision rule $p(D|V)$, consider the conditional loss $L(S)$ for all possible signals, and let $[L(S)]_{max}$ be the maximum value attained by $L(S)$. We seek that decision rule for which $[L(S)]_{max}$ is as small as possible. As we noted in Chapter 13, this criterion concentrates attention on the worst possible case regardless of the probability of occurrence of this case, and it is thus in a sense too conservative. It has, however, for some the psychological advantage that it does not involve the prior probabilities of the different signals, $p(S|X)$, and therefore it can be applied by persons who, under the handicap of orthodox training, have a mental hangup against prior probabilities.

The Bayes Criterion. We seek that decision rule for which the average loss $\langle L \rangle$ is minimized. In order to apply this, a prior distribution $p(S|X)$ must be available.

Other criteria were proposed before the days of Wald's decision theory. In the Neyman–Pearson theory, we fix the probability of occurrence of one type of error at some small value ϵ, and then minimize the probability δ of the other type of error subject to this constraint.† Arnold Siegert's "Ideal Observer" minimizes the total probability of error $(\epsilon + \delta)$.

† For example, we suspect that at an Early Warning Radar Installation, the primary constraint might be that the Commanding Officer shall not be roused out of bed by a false alarm more often than once per month, and subject to that requirement, we minimize the probablity of a false rest.

After having invented many different such *ad hoc* criteria from various viewpoints, and arguing their relative merits on philosophical grounds, the basic mathematical identity of all these criteria came as quite a surprise to the early workers in this field. We shall see below that all of them are special cases of the Bayes criterion.

Let us find the Bayes solution, as it was rationalized in decision theory. Substituting in succession equations (14–15), (14–6), and (14–3) into (14–16), we obtain for the average loss

$$\langle L\rangle = \sum_{DV}\left[\sum_S L(D,S)\,p(VS|X)\right]\,p(D|V). \tag{14–17}$$

If $L(D,S)$ is a definite function independent of $p(D|V)$ (this assumption excludes for the moment the information loss function), there is no function $p(D|V)$ for which this expression is stationary in the sense of calculus of variations. We then minimize $\langle L\rangle$ merely by choosing for each possible V that decision $D_1(V)$ for which

$$K(D_1,S) \equiv \sum_S L(D_1,S)\,p(VS|X) \tag{14–18}$$

is a minimum. Thus, we adopt the decision rule

$$p(D|V) = \delta(D,D_1). \tag{14–19}$$

In general there will be only one such D_1, and the best decision rule is nonrandom. However, in case of "degeneracy," $K(D_1,V) = K(D_2,V)$, any randomized rule of the form

$$p(D|V) = a\delta(D,D_1) + b\delta(D,D_2), \quad a+b=1 \tag{14–20}$$

is just as good by the criterion being used. This degeneracy occurs at "threshold" values of V, where we change from one decision to another.

A Discrete Example

Consider the case already mentioned, where there are two possible signals S_0, and S_1, and a loss matrix

$$L_{ij} = \begin{pmatrix} L_{00} & L_{10} \\ L_{10} & L_{11}\end{pmatrix} = \begin{pmatrix} 0 & L_r \\ L_a & 0\end{pmatrix}$$

where L_s, L_r are the losses incurred by a false alarm and a false rest, respectively. Then

$$\begin{aligned} K(D_0,V) &= L_{01}\,p(VS_1|X) = L_r\,p(VS_1|X) \\ K(D_1,V) &= L_{10}\,p(VS_0|X) = L_a\,p(VS_0|X)\end{aligned} \tag{14–21}$$

and the decision rule that minimizes $\langle L\rangle$ is

$$\begin{aligned} &\text{Choose } D_1 \text{ if} \quad \frac{p(VS_1|X)}{p(VS_0|X)} > \frac{L_a}{L_r} \\ &\text{Choose } D_0 \text{ if} \quad \frac{p(VS_1|X)}{p(VS_0|X)} < \frac{L_a}{L_r} \\ &\text{Choose either at random in case of equality.}\end{aligned} \tag{14–22}$$

If the prior probabilities of signal and no signal are

$$p(S_1|X) = p, \qquad p(S_0|X) = q = 1-p \tag{14–23}$$

respectively, the decision rule becomes

$$\text{Choose } D_1 \text{ if} \qquad \frac{p(V|S_1)}{p(V|S_0)} > \frac{qL_a}{pL_r}, \quad etc. \tag{14–24}$$

The left–hand side of (14–24) is a likelihood ratio, which depends only on the *pdf* assigned to the noise, and is the quantity which should be computed by the optimum receiver according to the Bayes criterion.

This same quantity is essential one regardless of the assumed loss function and regardless of the probability of occurrence of the signal; these affect only the threshold of detection. Furthermore, if the receiver merely computes this likelihood ratio and delivers it at the output without making any decision, it provides us with all the information we need to make optimum decisions in the Bayes sense. Note particularly the generality of this result, which is one of the most important ones for our applications; no assumptions are needed as to the type of signal, linearity of the system, or properties of the noise.

We now work out, for purposes of illustration, the decision rules and their degree of reliability, for several of the above criteria, in the simplest possible problem. We have a linear system in which the voltage is observed at a single instant, and we are to decide whether a signal, which can have only amplitude S_1, is present in noise, to which we assign a gaussian *pdf* with mean square value σ^2:

$$W(N) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{N^2}{2\sigma^2}\right]. \tag{14–25}$$

The likelihood ratio in (14–24) then becomes

$$\frac{p(V|S_1)}{p(V|S_0)} = \frac{W(V-S_1)}{W(V)} = \exp\left[\frac{2VS_1 - S_1^2}{2\langle\sigma^2\rangle}\right] \tag{14–26}$$

and since this is a monotonic function of V, the decision rule can be written as

$$\text{Choose } \begin{pmatrix} D_1 \\ D_0 \end{pmatrix} \quad \text{when} \quad V \begin{pmatrix} > \\ < \end{pmatrix} V_b \tag{14–27}$$

with

$$\frac{V_b}{\sigma} = \frac{1}{2s}\left[2\log\left(\frac{qL_a}{pL_r}\right) + s^2\right] = v_b \tag{14–28}$$

in which

$$s \equiv \frac{S_1}{\sigma} \text{ is the voltage signal-to-noise ratio, and}$$

$$v \equiv \frac{V}{\sigma} \text{ is the normalized voltage.}$$

Now we find for the probability of a false rest:

$$\begin{aligned} p(R|X) = p(D_0S_1|X) &= p\sum_V p(D_0|V)\,p(V|S_1) = p\int_{-\infty}^{V_b} dV W(V-S_1) \\ &= p\Phi(v_b - s) \end{aligned} \tag{14–29}$$

and for a false alarm,

$$p(A|X) = p(D_1 S_0|X) = q \sum_V p(D_1|V)\, p(V|S_0) = q \int_{V_b}^{\infty} dV W(V) \\ = q[1 - \Phi(v_b)]. \tag{14–30}$$

Here $\Phi(x)$ is the cumulative normal distribution function, calculated and tabulated in Appendix I:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt\,. \tag{14–31}$$

For $x > 2$, a good approximation is

$$1 - \Phi(x) \approx \frac{e^{-x^2/2}}{x\sqrt{2\pi}} \tag{14–32}$$

As a numerical example, if $L_r = 10L_a$, $q = 10p$, these expressions reduce to

$$p(A|X) = 10p(R|X) = \frac{10}{11}\left[1 - \Phi\left(\frac{s}{2}\right)\right]\,. \tag{14–33}$$

The probability of a false alarm is less than 0.027, and of a false rest less than 0.0027 for $s > 4$. For $s > 6$, these numbers become 1.48×10^{-3}, 1.48×10^{-4} respectively.

Let us see what the minimax criterion would give in this problem. The conditional losses are

$$L(S_0) = L_a \sum_V p(D_1|V)\, p(V|S_0) = L_a \int_{-\infty}^{\infty} p(D_1|V)\, W(V) dV \\ L(S_1) = L_r \sum_V p(D_0|V)\, p(V|S_1) = L_r \int_{-\infty}^{\infty} p(D_0|V)\, W(V - S_1) dV \tag{14–34}$$

Writing $f(V) \equiv p(D_1|V) = 1 - p(D_0|V)$, the only restriction on $f(V)$ is $0 \le f(V) \le 1$. Since L_a, L_r, and $W(V)$ are all positive, a change $\delta f(V)$ in the neighborhood of any given point V will always increase one of the quantities (14–34) and decrease the other. Thus when the maximum $L(S)$ has been made as small as possible, we will certainly have $L(S_0) = L(S_1)$, and the problem is thus to minimize $L(S_0)$ subject to this constraint.

Suppose that for some particular $p(S|X)$ the Bayes decision rule happened to give $L(S_0) = L(S_1)$. Then this particular solution must be identical with the minimax solution, for with the above constraint, $\langle L \rangle = [L(S)]_{max}$, and if the Bayes solution minimizes $\langle L \rangle$ with respect to all variations $\delta f(V)$ in the decision rule, it *a fortiori* minimizes it with respect to the smaller class of variations which keep $L(S_0) = L(S_1)$. Therefore the decision rule will have the same form as before: There is a minimax threshold V_m such that

$$f(V) = \left\{ \begin{array}{ll} 0, & V < V_m \\ 1, & V > V_m \end{array} \right\}\,. \tag{14–36}$$

Any change in V_m from the value which makes $L(S_0) = L(S_1)$ necessarily increases one or the other of these quantities. The equation determining V_m is therefore

$$L_a \int_{V_m}^{\infty} W(V) dV = L_r \int_{-\infty}^{V_m} W(V - S_1) dV$$

or, in terms of normalized quantities,

$$L_a[1-\Phi(v_m)] = L_r\Phi(v_m - s). \tag{14–37}$$

Note that (14–30), (14–31) give the conditional probabilities of false rest and false alarm for any decision rule of type (14–36), regardless of whether the threshold was determined from (14–28) or not; for the arbitrary threshold V_0

$$\begin{aligned} p(R|S_1) &= p(V < V_0|S_1) = \Phi(v_0 - s) \\ p(A|S_0) &= p(V > V_0|S_0) = \frac{1}{2}[1-\Phi(v_0)]. \end{aligned} \tag{14–38}$$

From (14–28) we see that there is always a particular ratio (p/q) which makes the Bayes threshold V_b equal to the minimax threshold V_m. For values of (p/q) other than this worst value, the Bayes criterion gives a lower average loss than does the minimax, although one of the conditional losses $L(S_0)$, $L(S_1)$ will be greater than the minimax value.

These relations and several previous remarks are illustrated in Figure 14.1, in which we plot the conditional losses $L(S_0)$, $L(S_1)$ and the average loss $\langle L\rangle$ as functions of the threshold V_0, for the case $L_a = (3/2)L_r$, $p = q = 1/2$. The minimax threshold is at the common crossing–point of these curves, while the Bayes threshold occurs at the lowest point of the $\langle L\rangle$ curve.

One sees how the Bayes threshold moves as the ratio (p/q) is varied, and in particular that the value of (p/q) which makes $V_b = V_m$ also leads to the maximum value of the $\langle L\rangle_{min}$ obtained by the Bayes criterion. Thus we could also define a "maximin" criterion; first find the Bayes decision rule which gives minimum $\langle L\rangle$ for a given $p(S|X)$, then vary the prior probability $p(S|X)$ until the maximum value of $\langle L\rangle_{min}$ is attained. The decision rule thus obtained is identical with the one resulting from the minimax criterion; this is the worst possible prior probability, in the sense that the most pessimistic rule is the best that can be done.

The Neyman–Pearson criterion is easily discussed in this example: Suppose the conditional probability of a false alarm $p(D_1|S_0)$ is held fixed at some value ϵ, and we wish to minimize the conditional probability $p(D_0|S_1)$ of a false rest, subject to this constraint. Now the Bayes criterion minimizes the average loss

$$\langle L\rangle = pL_r(D_0|S_1) + qL_r(D_1|S_0)$$

with respect to any variation $\delta p(D|V)$ in the decision rule. In particular, therefore, it minimizes it with respect to the smaller class of variations which hold $p(D_1|S_0)$ constant at the value finally obtained. Thus it minimizes $p(D_0|S_1)$ with respect to these variations and solves the Neyman–Pearson problem; we need only choose the particular value of the ratio (qL_a/pL_r) which results in the assumed value of ϵ according to (14–28), (14–30).

We find for the Neyman-Pearson threshold, from (14–38)

$$\Phi(v_{np}) = 1-\epsilon \tag{14–39}$$

and the conditional probability of detection is

$$p(D_1|S_1) = 1 - p(D_0|S_1) = \Phi(s - v_{np}). \tag{14–40}$$

If $\epsilon = 10^{-3}$, a detection probability of 99 percent or better is attained for $s > 6$.

It is important to note that these numerical examples depend critically on our noise *pdf* assignment. If we have prior information about the noise beyond its first and second moments, the noise *pdf* expressing this may not be gaussian, and the actual situation may be either more or less favorable than indicated by the above relations.

It is well known that in one sense noise with a gaussian frequency distribution is the worst possible kind; because of its maximum entropy properties, it can obscure a weak signal more completely than can any other noise of the same average power. On the other hand, gaussian noise is a very favorable kind from which to extract a fairly strong signal, because the probability that the noise will exceed a few times the RMS value $\sigma = \sqrt{\langle N^2 \rangle}$ becomes vanishing small. Consequently, the probability of making an incorrect decision on the presence or absence of a signal goes to zero very rapidly as the signal strength is increased. The high reliability of operation found above for $s > 6$ would not be found for noise possessing a probability distribution with wider tails.

The type of noise frequency distribution to be expected in any particular case depends, of course, on the physical mechanism which gives rise to the noise. When the noise is the result of a large number of small, independent effects, the Landon derivation of Chapter 7 and the Central Limit Theorem both tell us that a gaussian frequency distribution for the total noise is by far the most likely to be found, regardless of the nature of the individual sources.

All of these apparently different decision criteria lead, when worked out, to a probability ratio test. In the case of a binary decision, it took the simple form (14–22). Of course, any decision process can be broken down into successive binary decision, so this case really has the whole story in it. All the different criteria amounted, in the final analysis, only to different philosophies about how you choose the threshold value at which you change your decision.

How Would Our Robot Do It?

Now let's see how this problem appears from the viewpoint of our robot. The rather long arguments we had to go through above (and even they are very highly condensed from the original literature) to get the result are due only to the orthodox view which insists on looking at the problem backwards, *i.e.*, on concentrating attention on the final decision rather than on the inference process which logically has to precede it.

To the robot, if our job is to make the best possible decision as to whether the signal is present, the obvious thing we must do is to calculate the *probability* that the signal is present, conditional on all the evidence at hand. If there are only two possibilities, S_0, S_1, to be taken into account, then after we have seen voltage V, the posterior odds on S_1 are from (4–7),

$$O(S_1|VX) = O(S_1|X)\,\frac{p(V|S_1)}{p(V|S_0)}. \tag{14–41}$$

If we give the robot the loss function (14–21) and ask it to make the decision which minimizes the expected loss, it will evidently use the decision rule

$$\text{choose } D_1 \text{ if } O(S_1|V) = \frac{p(S_1|V)}{p(S_0|V)} > \frac{L_a}{L_r}\,, \tag{14–42}$$

etc. But from the product rule, $p(VS_1|X) = p(S_1|V)\,p(V|X)$, $p(VS_0|X) = p(S_0|V)\,p(V|X)$, and (14–42) is identical with (14–22). So, just from looking at this problem the other way around, our robot obtains the same final result in just two lines!

You see that all this discussion of strategies, admissibility, conditional losses, etc., was completely unnecessary. Except for the introduction of the loss function at the end, there is nothing in the actual functional operation of Wald's decision theory that isn't contained already in basic *probability* theory, if we will only use it in the full generality given to it by James Bernoulli and Laplace.

Historical Remarks

This comparison shows why the development of decision theory has, more than any other single factor, provoked our 'Bayesian Revolution' in statistical thought. For some fifty years, Harold Jeffreys tried valiantly to explain the great advantages of the Laplace methods to statisticians, and his efforts met only with a steady torrent of denials and ridicule. It was then a real irony that the work of one of the most respected of "orthodox" statisticians (Abraham Wald), which was hailed, very properly, as a great advance in statistical practice, turned out to give, after very long and complicated arguments, exactly the same final results that the despised Laplace methods give you immediately. Wald showed in great generality what we have just illustrated by one simple example.

The only proper conclusion, as a few recognized at once, is that the supposed distinction between statistical inference and probability theory was entirely artificial – a tragic error of judgment which has wasted perhaps a thousand man–years of our best mathematical talent in the pursuit of false goals.

In the works cited, addressed to electrical engineers, the viewpoint of Middleton and van Meter was that of the Neyman–Pearson and Wald decision theories. At about the same time, Herbert Simon expounded the Neyman–Pearson viewpoint to economists. The writer collaborated with David Middleton for a short time while he was writing his large work, and tried to persuade him of the superiority of the straight Bayesian approach to decision theory. The success of the effort may be judged by comparing Middleton's Chapter 18 – particularly its length – with our exposition. It seems that persons with orthodox training had received such strong anti–Bayesian indoctrination that we were in an infinite regress situation; although they could not deny the results that Bayesians got on any specific problem, they could never believe that Bayesian methods would work on the next problem until that next solution was also presented to them.

The Classical Matched Filter A funny thing happened in the history of this subject. In the 1930's, Electrical Engineers knew nothing whatsoever about probability theory; they knew about signal to noise ratios. Receiver input circuits were designed for many years on the basis that signal to noise ratio was maximized by empirical trial and error. Then a general theoretical result was found: if you take the ratio of (peak signal)^2 to mean square noise, and find, as a variational principle, the design of input stages of the receiver which will maximize it, this turned out to have an analytically neat and useful solution. It is now called the *classical matched filter*, and it has been discovered independently by dozens of people.

To the best of our knowledge, the first person to derive this matched filter solution was the late Professor W. W. Hansen of Stanford University. The writer was working with him, beginning in May 1942, on problems of radar detection. Shortly before then, Hansen had circulated a little memorandum dated 1941, in which he gave this solution for the design of the optimum response curve for the receiver first stage. Years later I was thinking about an entirely different problem (an optimum antenna pattern for a radar system to maximize the ratio (signal)/(ground clutter response), and when I finally got the solution, I recognized it as the same result that Bill Hansen had shown me many years before. This theory is considered later, under "Optimum Antenna and Filter Design."

Throughout the 1950's, almost every time one opened a journal concerned with these problems, somebody else had a paper announcing the discovery of the same solution. The situation was satirized in a famous editorial by Peter Elias (1958), entitled "Three Famous Papers". He suggested that it was high time that people stopped rediscovering the easiest solution, and started to think about the many harder problems still in need of solution.

But also, in the 1950's people got more sophisticated about the way they handled their detection problems, and they started using this wonderful new tool, statistical decision theory, to see if there were still better ways of handling these design problems. The strange thing happened that in the case of a linear system with gaussian noise, the optimum solution which decision theory leads you to, turns out to be exactly the same old classical matched filter. At first glance, it was very surprising that two approaches so entirely different conceptually should lead to the same solution. But, note that our robot represents a viewpoint from which it is obvious that the two lines of argument would have to give the same result.

To our robot it is obvious that the best analysis you can make of the problem will always be one in which you calculate the *probabilities* that the various signals are present by means of Bayes' theorem (but to those with orthodox training this was not obvious; it was vehemently denied). But let us apply Bayes' theorem in the Logarithmic form of Chapter 4. If we now let S_0 and S_1 stand for numerical values giving the amplitude of two possible signals, as a function of V the *evidence* for S_1 is increased by

$$\log \frac{p(V|S_1)}{p(V|S_0)} = \frac{(V-S_0)^2 - (V-S_1)^2}{2\langle\sigma^2\rangle} = \text{ const. } + \frac{(S_1 - S_0)}{\langle\sigma^2\rangle} V \,. \tag{14–43}$$

In the case of a linear system with gaussian noise, the observed voltage is itself just a linear function of the posterior probability measured in db. So, they are essentially just two different ways of formulating the same problem. Without recognizing it, we had essentially solved this problem already in the Bayesian hypothesis testing discussion of Chapter 4.

In England, P. M. Woodword had perceived much of this correctly in the 1940's – but he was many years ahead of his time. Those with conventional statistical training were unable to see any merit in his work, and simply ignored it. His book (Woodword, 1953) is highly recommended reading; although it does not solve any of our current problems, its thinking is still in advance of some current literature and practice.

We have seen that the other non–Bayesian approaches to the theory all amounted to different philosophies of how you choose the threshold at which you change your decision. Because of the fact that they all lead to the same probability ratio test, they must necessarily all be derivable from Bayes' theorem.

The problem just examined by several different decision criteria is, of course, the simplest possible one. In a more realistic problem we will observe the voltage $V(t)$ as a function time, perhaps several voltages $V_1(t)$, $V_2(t), \cdots$ in several different channels. We may have many different possible signals $S_a(t)$, $S_b(t), \cdots$ to distinguish and corresponding many possible decisions. We may need to decide not only *whether* a given signal is present, but also to make the best estimates of one or more signal parameters (such as intensity, starting time, frequency, phase, rate of frequency modulation, *etc.*). Therefore, just as in the problem of quality control discussed in Chapter 4, the details can become arbitrarily complicated. But these extension are, from the Bayesian viewpoint, straightforward in that they require no new principles beyond those already given, only mathematical generalization.

We shall return to some of these more complicated problems of detection and filtering when we take up frequency/shape estimation; but for now let's look at another elementary kind of decision problem. In the ones just discussed, we needed Bayes' theorem, but not maximum entropy. Now we examine a kind of decision problem where we need maximum entropy, but not Bayes' theorem.

The Widget Problem

This problem was first propounded at a symposium held at Purdue University in November, 1960 – at which time, however, the full solution was not known. This was worked out later (Jaynes, 1963c), and some numerical approximations were improved in the computer work of Tribus and Fitts (1968).

The widget problem has proved to be interesting in more respects than originally realized. It is a decision problem in which there is no occasion to use Bayes' theorem, because no "new" information is acquired. Thus it would be termed a "no data" decision problem in the sense of Chernoff and Moses (1959). However, at successive stages of the problem we have more and more prior information; and digesting it by maximum entropy leads to a sequence of prior probability assignments, which lead to different decisions. Thus it is an example of the "pure" use of maximum entropy, as in statistical mechanics. It is hard to see how the problem could be formulated mathematically at all without use of maximum entropy, or some other device [like the method of Darwin & Fowler (1928) in Statistical Mechanics, or the 'method of the most probable distribution' dating back to Boltzmann (1871)] which turns out in the end to be mathematically equivalent to maximum entropy.

The problem is interesting also in that we can see a continuous gradation from decision problems so simple that common sense tells us the answer instantly with no need for any mathematical theory, through problems more and more involved so that common sense has more and more difficulty in making a decision, until finally we reach a point where nobody has yet claimed to be able to see the right decision intuitively, and we require the mathematics to tell us what to do.

Finally, it turns out to be very close to an important real problem faced by oil prospectors. The details of the real problem are shrouded in proprietary caution; but it is not giving away any secrets to report that, a few years ago, the writer spent a week at the research laboratories of one of our large oil companies, lecturing for over 20 hours on the widget problem. We went through every part of the calculation in excruciating detail – with a room full of engineers armed with calculators, checking up on every stage of the numerical work.

Here is the problem: Mr. A is in charge of a Widget factory, which proudly advertises that it can make delivery in 24 hours on any size order. This, of course, is not really true, and Mr. A's job is to protect, as best he can, the Advertising Manager's reputation for veracity. This means that each morning he must decide whether the day's run of 200 widgets will be painted red, yellow or green. (For complex technological reasons, not relevant to the present problem, only one color can be produced per day.) We follow his problem of decision through several stages of increasing knowledge.

Stage 1. When he arrives at work, Mr. A checks with the stock room and finds that they now have in stock 100 red widgets, 150 yellow, and 50 green. His ignorance lies in the fact that he does not know how many orders for each type will come in during the day. Clearly, in this state of ignorance, Mr. A will attach the highest significance to any tiny scrap of information about orders likely to come in today; and if no such scraps are to be had, we do not envy Mr. A his job. Still, if a decision must be made here and now on no more information that this, his common sense will probably tell him that he had better build up that stock of green widgets.

Stage 2. Mr. A, feeling the need for more information, calls up the front office and asks, "Can you give me some idea of how many orders for red, yellow, and green widgets are likely to come in today?" They reply, "Well, we don't have the breakdown of what has been happening each day, and it would take us a week to compile that information from our files. But we do have a summary of the total sales last year. Over the last year, we sold a total of 13,000 red, 26,000 yellow, and 2600 green. Figuring 260 working days, this means that last year we sold an average of 50 red, 100 yellow, and 10 green each day." If Mr. A ponders this new information for a few seconds I think he will change his mind, and decide to make yellow ones today.

Stage 3. The man in the front office calls Mr. A back and says, "It just occurred to me that we do have a little more information that might possibly help you. We have at hand not only the total number of widgets sold last year, but also the total number of orders we processed. Last year we got a total of 173 orders for red, 2600 for yellow, and 130 for green. This means that the customers who use red widgets order, on the average, $13000/173 = 75$ widgets per order, while the average order for yellow and green were $26000/2600=10$, and $2600/130=20$ respectively." These new data do not change the expected daily demand; but if Mr. A is very shrewd and ponders it very hard, I think he may change his mind again, and decide to make red ones today.

Stage 4. Mr. A is just about to give the order to make red widgets when the front office calls him again to say, "We just got word that a messenger is on his way here with an emergency order for 40 green widgets." Now, what should he do? Up to this point, Mr. A's decision problem has been simple enough so that reasonably good common sense will tell him what to do. But now, he is in trouble; qualitative common sense is just not powerful enough to solve his problem, and he needs a mathematical theory to determine a definite optimum decision.

Let's summarize all the above data in a table:

Table 14.1 Summary of Four stages of the Widget Problem

	R	Y	G	Decision
1. In stock	100	150	50	G
2. Avg. Daily Order Total	50	100	10	Y
3. Avg. Individual Order	75	10	20	R
4. Specific Order			40	?

In the last column we give the decision that seemed intuitively to be the best ones before we had worked out the mathematics. Do other people agree with this intuitive judgment? Professor Myron Tribus has put this to a test by giving talks about this problem, and taking votes from the audience before the solution is given. We quote his findings as given in their paper (M. Tribus and G. Fitts, 1968). They use D_1, D_2, D_3, D_4 to stand for the optimum decisions in stages 1, 2, 3, 4 respectively:

"Before taking up the formal solution, it may be reported that Jaynes' widget problem has been presented to many gatherings of engineers who have been asked to vote on D_1, D_2, D_3, D_4. There is almost unanimous agreement about D_1. There is about 85 percent agreement on D_2. There is about 70 percent agreement on D_3, and almost no agreement on D_4. One conclusion stands out from these informal tests; the average engineer has remarkably good intuition in problems of this kind. The majority vote for D_1, D_2, and D_3 has always been in agreement with the formal mathematical solution. However, there has been almost universal disagreement over how to defend the intuitive solution. That is, while many engineers could agree on the best course of action, they were much less in agreement on *why* that course was the best one."

Solution For Stage 2

Now, how are we to set up this problem mathematically? In a real life situation, evidently, the problem would be a little more complicated than indicated so far, because what Mr. A does today also affects how serious his problem will be tomorrow. That would get us into the subject of dynamic programming. But for now, just to keep the problem simple, we shall solve only the truncated problem in which he makes decisions on a day to day basis with no thought of tomorrow.

We have just to carry out the steps enumerated under "General Decision Theory" at the end of the last Chapter. Since Stage 1 is almost too trivial to work with, consider the problem of Stage 2. First, we define our underlying hypothesis space by enumerating the possible "states of nature" θ_j that we will consider. These correspond to all possible order situations that could arise; if Mr. A knew in advance exactly how many red, yellow, and green widgets would be ordered today, his decision problem would be trivial. Let $n_1 = 0, 1, 2, \ldots$ be the number of red widgets that will be ordered today, and similarly n_2, n_3 for yellow and green respectively. Then any conceivable order situation is given by specifying three non–negative integers $\{n_1, n_2, n_3\}$. Conversely, every ordered triple of non–negative integers represents a conceivable order situation.

Next, we are to assign prior probabilities $p(\theta_j|X) = p(n_1 n_2 n_3|X)$ to the states of nature, which maximize the entropy of the distribution subject to the constraints of our prior knowledge. We solved this problem in general in Chapter 11, Equations (11–27)–(11–35); and so we just have to translate the result into our present notation. The index i on x_i in Chapter 11 now corresponds to the three integers n_1, n_2, n_3; the function $f_k(x_i)$ also corresponds to the n_i, since the prior information at this stage will be used to fix the expectations $\langle n_1\rangle, \langle n_2\rangle, \langle n_3\rangle$ of orders for red, yellow, and green widgets as 50, 100, 10 respectively. With three constraints we will have three Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3$, and the partition function (11–31) becomes

$$Z(\lambda_1,\lambda_2,\lambda_3) = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\sum_{n_3=0}^{\infty} \exp(-\lambda_1 n_1 - \lambda_2 n_2 - \lambda_3 n_3) = \prod_{i=1}^{3}(1-e^{-\lambda_j})^{-1}. \tag{14–44}$$

The λ_i are determined from (11–32):

$$\langle n_i\rangle = -\frac{\partial}{\partial\lambda_i}\log Z = \left(\frac{1}{e^{\lambda_i}-1}\right).$$

The maximum entropy probability assignment (11–28) for the states of nature $\theta_j = \{n_1 n_2 n_3\}$ therefore factors:

$$p(n_1 n_2 n_3) = p_1(n_1)p_2(n_2)p_3(n_3) \tag{14–46}$$

with

$$\begin{aligned} p_i(n_i) &= (1-e^{-\lambda_i})e^{-\lambda_i n_i}, \quad n_i = 1,2,3\ldots \\ &= \frac{1}{\langle n_i\rangle + 1}\left[\frac{\langle n_i\rangle}{\langle n_i\rangle+1}\right]^{n_i}. \end{aligned} \tag{14–47}$$

Thus in Stage 2, Mr. A's state of knowledge about today's orders is given by three exponential distributions:

$$p_1(n_1) = \frac{1}{51}\left(\frac{50}{51}\right)^{n_1}, \qquad p_2(n_2) = \frac{1}{101}\left(\frac{100}{101}\right)^{n_2}, \qquad p_3(n_3) = \frac{1}{11}\left(\frac{10}{11}\right)^{n_3}. \tag{14–48}$$

Applications of Bayes' theorem to digest new evidence E is absent because there is no new evidence. Therefore, the decision must be made directly from the prior probabilities (14–48), as is always the case in statistical mechanics.

So, we now proceed to enumerate the possible decisions. These are $D_1 \equiv$ make red ones today, $D_2 \equiv$ make yellow ones, $D_3 \equiv$ make green ones, for which we are to introduce a loss function $L(D_i, \theta_j)$. Mr. A's judgment is that there is no loss if all orders are filled today; otherwise the loss will be proportional to – and in view of the invariance of the decision rule under proper linear transformations that we noted at the end of Chapter 13, we may as well take it equal to – the total number of unfilled orders.

The present stock of red, yellow, and green widgets is $S_1 = 100$, $S_2 = 150$, $S_3 = 50$ respectively. On decision D_1 (make red widgets) the available stock S_1 will be increased by the day's run of 200 widgets, and the loss will be

$$L(D_1; n_1 n_2 n_3) = g(n_1 - S_1 - 200) + g(n_2 - S_2) + g(n_3 - S_3) \tag{14–49}$$

where $g(x)$ is the ramp function

$$g(x) \equiv \begin{cases} x, & x \geq 0 \\ 0, & x \leq 0 \end{cases}. \tag{14–50}$$

Likewise, on decisions D_2, D_3 the loss will be

$$L(D_2; n_1 n_2 n_3) = g(n_1 - S_1) + g(n_2 - S_2 - 200) + g(n_3 - S_3), \tag{14–51}$$

$$L(D_3; n_1 n_2 n_3) = g(n_1 - S_1) + g(n_2 - S_2) + g(n_3 - S_3 - 200). \tag{14–52}$$

So, if decision D_1 is made, the expected loss will be

$$\begin{aligned} \langle L \rangle_1 &= \sum_{N_i} p(n_1 n_2 n_3) L(D_1; n_1 n_2 n_3) \\ &= \sum_{n_1=0}^{\infty} p_1(n_1) g(n_1 - S_1 - 200) + \sum_{n_2=0}^{\infty} p_2(n_2) g(n_2 - S_2) + \sum_{n_3=0}^{\infty} p_3(n_3) g(n_3 - S_3) \end{aligned}$$

and similarly for D_2, D_3. The summations are elementary, giving

$$\begin{aligned} \langle L \rangle_1 &= \langle n_1 \rangle e^{-\lambda_1 (S_1 + 200)} + \langle n_2 \rangle e^{-\lambda_2 S_2} + \langle n_3 \rangle e^{-\lambda_3 S_3}, \\ \langle L \rangle_2 &= \langle n_1 \rangle e^{-\lambda_1 S_1} + \langle n_2 \rangle e^{-\lambda_2 (S_2 + 200)} + \langle n_3 \rangle e^{-\lambda_3 S_3}, \\ \langle L \rangle_3 &= \langle n_1 \rangle e^{-\lambda_1 S_1} + \langle n_2 \rangle e^{-\lambda_2 S_2} + \langle n_3 \rangle e^{-\lambda_3 (S_3 + 200)} \end{aligned} \tag{14–54}$$

or, inserting numerical values

$$\begin{aligned} \langle L \rangle_1 &= 0.131 + 22.48 + 0.085 = 22.70 \\ \langle L \rangle_2 &= 6.902 + 3.073 + 0.085 = 10.6 \\ \langle L \rangle_3 &= 6.902 + 22.48 + 4 \times 10^{-10} = 39.38 \end{aligned} \tag{14–55}$$

showing a strong preference for decision $D_2 \equiv$ "make yellow ones today," as common sense had already anticipated.

Physicists will recognize that Stage 2 of Mr. A's decision problem is mathematically the same as the theory of harmonic oscillators in quantum statistical mechanics. There is still another engineering application of the harmonic oscillator equations, in some problems of message encoding, to be noted when we take up communication theory. We are trying to emphasize the generality of this theory, which is mathematically quite old and well known, but which has been applied in the past only in some specialized problems in physics. This general applicability can be seen only after we are emancipated from the orthodox view that all probability distributions must be interpreted in the frequency sense.

Solution For Stage 3

In Stage 3 of Mr. A's problem we have some additional pieces of information given the average *individual* orders for red, yellow, and green widgets. To take account of this new information, we need to go down into a deeper hypothesis space; set up a more detailed enumeration of the states of nature in which we take into account not only the total orders for each type, but also the breakdown into individual orders. We could have done this also in Stage 2, but since at that stage there was no information available bearing on this breakdown, it would have added nothing to the problem (the subtle difference that this makes after all will be noted later).

In Stage 3, a possible state of nature can be described as follows. We receive u_1 individual orders for 1 red widget each, u_2 orders for 2 red widgets each, ..., u_r individual orders for r red widgets each. Also, we receive v_y orders for y yellow widgets each, and w_g orders for g green widgets each. Thus a state of nature is specified by an infinite number of non-negative integers

$$\theta = \{u_1 u_2 \ldots; v_1 v_2 \ldots; w_1 w_2 \ldots\} \tag{14–56}$$

and conversely every such set of integers represents a conceivable state of nature, to which we assign a probability $p(u_1 u_2 \ldots; v_1 v_2 \ldots; w_1 w_2 \ldots)$.

Today's total demands for red, yellow and green widgets are, respectively

$$n_1 = \sum_{r=1}^{\infty} r u_r, \qquad n_2 = \sum_{y=1}^{\infty} y v_y, \qquad n_3 = \sum_{g=1}^{\infty} g w_g,$$

the expectations of which were given in Stage 2 as $\langle n_1 \rangle = 50$, $\langle n_2 \rangle = 100$, $\langle n_3 \rangle = 10$. The total number of individual orders for red, yellow, and green widgets are respectively

$$m_1 = \sum_{r=1}^{\infty} u_r, \qquad m_2 = \sum_{y=1}^{\infty} v_y, \qquad m_3 = \sum_{g=1}^{\infty} w_g,$$

and the new feature of Stage 3 is that $\langle m_1 \rangle$, $\langle m_2 \rangle$, $\langle m_3 \rangle$ are also known. For example, the statement that the average individual order for red widgets is 75 means that $\langle n_1 \rangle = 75 \langle m_1 \rangle$.

With six average values given, we will have six Lagrange multipliers $\{\lambda_1 \mu_1; \lambda_2 \mu_2; \lambda_3 \mu_3\}$. The maximum entropy probability assignment will have the form

$$p(u_1 u_2 \ldots; v_1 v_2 \ldots; w_1 w_2 \ldots) = \exp(-\lambda_0 - \lambda_1 n_1 - \mu_1 m_1 - \lambda_2 n_2 - \mu_2 m_2 - \lambda_3 n_3 - \mu_3 m_3)$$

which factors:

$$p(u_1 u_2 \ldots; v_1 v_2 \ldots; w_1 w_2 \ldots) = p_1(u_1 u_2 \ldots) p_2(v_1 v_2 \ldots) p_3(w_1 w_2 \ldots). \tag{14–59}$$

The partition function also factors:

$$Z = Z_1(\lambda_1 \mu_1) Z_2(\lambda_2 \mu_2) Z(\lambda_3 \mu_3) \tag{14–60}$$

with

$$\begin{aligned} Z_1(\lambda_1 \mu_1) &= \sum_{u_1=1}^{\infty} \sum_{u_2=1}^{\infty} \cdots \exp[-\lambda_1(u_1 + 2u_2 + 3u_3 + \ldots) - \mu_1(u_1 + u_2 + u_3 + \ldots)] \\ &= \prod_{r=1}^{\infty} \frac{1}{1 - e^{-r\lambda_1 - \mu}} \end{aligned} \tag{14–61}$$

and similar expressions for Z_2, Z_3. To find λ_1, μ_1 we apply the general rule, Eq. (10–32):

$$\langle n_1 \rangle = \frac{\partial}{\partial \lambda_1} \sum_{r=1}^{\infty} \log(1 - e^{-r\lambda_1 - \mu_1}) = \sum_{r=1}^{\infty} \frac{r}{e^{r\lambda_1 + \mu_1} - 1}, \tag{14–62}$$

$$\langle m_1 \rangle = \frac{\partial}{\partial \mu_1} \sum_{r=1}^{\infty} \log(1 - e^{-r\lambda_1 - \mu_1}) = \sum_{r=1}^{\infty} \frac{1}{e^{r\lambda_1 + \mu_1} - 1}. \tag{14–63}$$

Combining with Eqs. (14–57), (14–58), we see that

$$\langle u_r \rangle = \frac{1}{e^{r\lambda_1 + \mu_1} - 1} \tag{14–64}$$

and now the secret is out – stage 3 of Mr. A's decision problem is just the theory of the ideal Bose–Einstein gas in quantum statistical mechanics!

If we treat the ideal Bose–Einstein gas by the method of the Gibbs grand canonical ensemble, we obtain just these equations, in which the number r corresponds to the r'th single–particle energy level, u_r to the number of particles in the r'th state, λ_1 and μ_1 to the temperature and chemical potential.

In the present problem it is clear that for all r, $\langle u_r \rangle \ll 1$, and that $\langle u_r \rangle$ cannot decrease appreciably below $\langle u_1 \rangle$ until r is of the order of 75, the average individual order. Therefore, μ_1 will be numerically large, and λ_1 numerically small, compared to unity. This means that the series (14–62), (14–63) converge very slowly and are useless for numerical work unless you write a computer program to do it. However, we can do it analytically if we transform them into rapidly converging sums as follows:

$$\sum_{r=1}^{\infty} \frac{1}{e^{\lambda r + \mu} - 1} = \sum_{r=1}^{\infty} \sum_{n=1}^{\infty} e^{-n(\lambda r + \mu)} = \sum_{n=1}^{\infty} \frac{e^{-n\mu}}{1 - e^{-n\lambda}}. \tag{14–65}$$

The first term is already an excellent approximation. Similarly,

$$\sum_{r=1}^{\infty} \frac{r}{e^{\lambda r + \mu} - 1} = \sum_{n=1}^{\infty} \frac{e^{-n(\lambda r + \mu)}}{(1 - e^{-n\lambda})^2} \tag{14–66}$$

and so (14–62) and (14–63) become

$$\langle n_1 \rangle = \frac{e^{-\mu_1}}{\lambda_1^2} \tag{14–67}$$

$$\langle m_1 \rangle = \frac{e^{-\mu_1}}{\lambda_1} \tag{14–68}$$

$$\lambda_1 = \frac{\langle m_1 \rangle_1}{\langle n_1 \rangle} = \frac{1}{75} = 0.0133, \tag{14–69}$$

$$e^{\mu_1} = \frac{\langle n_1 \rangle_1}{\langle m_1 \rangle} = 112.5, \tag{14–70}$$

$$\mu_1 = 4.722. \tag{14–71}$$

Tribus and Fitts, evaluating the sums by computer, get $\lambda_1 = 0.0131$, $\mu_1 = 4.727$; so our approximations (14–67), (14–68) are very good, at least in the case of red widgets.

The probability that u_r has a particular value is, from (14–59) or (14–61),

$$p(u_r) = (1 - e^{-r\lambda_1 - \mu})e^{(-r\lambda_1 + \mu_1)u_r} \tag{14–72}$$

which has the mean value (14–64) and the variance

$$\mathrm{var}(u_r) = \langle u_r^2 \rangle - \langle u_r \rangle^2 = \frac{e^{r\lambda_1+\mu_1}}{e^{r\lambda_1+\mu_1} - 1}\,. \tag{14–73}$$

The total demand for red widgets

$$n_1 = \sum_{r=1}^{\infty} r u_r \tag{14–74}$$

is expressed as the sum of a large number of independent terms. The *pdf* for n_1 will have the mean value (14–67) and the variance

$$\mathrm{var}(n_1) = \sum_{r=1}^{\infty} r^2 \mathrm{var}(u_r) = \sum_{r=1}^{\infty} \frac{r^2 e^{r\lambda_1+\mu_1}}{(e^{r\lambda_1+\mu_1} - 1)^2} \tag{14–75}$$

which we convert into the rapidly convergent sum

$$\sum_{r,n=1}^{\infty} n r^2 e^{-n(r\lambda+\mu)} = \sum_{n=1}^{\infty} n \frac{e^{-n(\lambda+\mu)} + e^{-n(2\lambda+\mu)}}{(1-e^{-n\lambda})^3} \tag{14–76}$$

or, approximately

$$\mathrm{var}(n_1) = \frac{2e^{-\mu_1}}{\lambda_1^3} = \frac{2}{\lambda_1}\langle n_1 \rangle. \tag{14–77}$$

At this point we can use some mathematical facts concerning the Central Limit Theorem. Because n_1 is the sum of a large number of small terms to which we have assigned independent probabilities, our probability distribution for n_1 will be very nearly gaussian:

$$p(n_1) \approx A \exp\left\{ -\frac{\lambda_1 (n_1 - \langle n_1 \rangle)^2}{4\langle n_1 \rangle} \right\} \tag{14–78}$$

for those values of n_1 which can arise in many different ways. For example, the case $n = 2$ can arise in only two ways: $u_1 = 2$, or $u_2 = 1$, all others u_k being zero. On the other hand, the case $n_1 = 150$ can arise in an enormous number of different ways, and the "smoothing" mechanism of the central limit theorem can operate. Thus, Eq. (14–78) will be a good approximation for the large values of n_1 of interest to us, but not for small n_1.

Then how accurately can Mr. A predict today's orders n_1 for red widgets? The (mean) $\pm$ (standard deviation) estimate from (14–78) is

$$(n_1)_{est} = \langle n_1 \rangle \pm \sqrt{\frac{2\langle n_1 \rangle}{\lambda_1}} = 50 \pm 86.6 \tag{14–79}$$

It was apparent from the start that his information is too meager to determine n_1 to any accuracy; yet the distribution does place a useful upper bound on the probable value. But this is a case where the probability distribution is so broad and skewed that the (mean) $\pm$ (standard deviation) is not a good criterion. The quartiles of (14–78) would tell us something more useful.