An Easy Derivation of Logistic Regression from the Bayesian and Maximum Entropy Perspective

Cite as: AIP Conference Proceedings **707**, 30 (2004); https://doi.org/10.1063/1.1751354
Published Online: 08 June 2004

David J. Blower

An Easy Derivation of Logistic Regression from the Bayesian and Maximum Entropy Perspective

David J. Blower

Institute for Human and Machine Cognition
40 S. Alcaniz, Pensacola, FL 32508
dblower@ihmc.us

Abstract. Logistic regression is a popular data analytic technique, but a compelling rationale for the equations that appear is missing in the conventional explanations. However, if one approaches logistic regression from a combined Bayesian and Maximum Entropy viewpoint, the explanation of its origin is relatively simple and direct. The perspective given here proceeds in two major steps. First, formally manipulate the probability symbols to rearrange them into the desired format. Here we want the probability of a binary criterion variable conditioned on knowledge of some number of predictor variables. The formal manipulation is done by using the sum and product rules of probability theory. Second, assign the numerical values to the joint probabilities that appear in Bayes's theorem by inserting information via Jaynes's Maximum Entropy Principle. The logistic regression equation then appears after a few more simple steps. A detailed numerical example is given to show the correspondence between the derivation given here and the conventional results. This application is just one example of scientific inference following the directions given by Jaynes in his *Probability Theory: The Logic of Science*.

INTRODUCTION

Logistic regression is a data analysis technique widely used by social scientists, economists, and medical researchers [1]. It also appears in a somewhat opaque fashion in the explanation of neural networks [2] and nonlinear classification [3]. Logistic regression is presented as a way to find the probability of some binary variable as a function of some number of predictor variables. These predictor variables may be continuous or discrete in nature.

For example, the probability of the presence or absence of coronary heart disease (CHD) might be modeled by a logistic regression equation as

$$P(\text{CHD}) = \frac{1}{1+\exp(-Y)} \tag{1}$$

where Y is a regression equation containing the regression parameters and the predictor variables. The regression equation would look like

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_q X_q \tag{2}$$

where the β_i are the regression parameters and the X_i are the coding variables for the discrete values taken on by the predictor variables. For the sake of an easy illustrative example, let's suppose that there are two predictor variables, a result from a laboratory

CP707, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23rd International Workshop,* edited by G. Erickson and Y. Zhai
2004 American Institute of Physics 0-7354-0182-9/04/$22.00

test and the patient's age. In the standard treatment, no rationale is given as to why the particular distribution in Equation (1) should be used and no explanation is proffered as to its origin. As is typical of the conventional approach, Equations (1) and (2) are presented out of the blue as a *fait accompli* unsupported by any underlying fundamental principles.

THE COMBINED BAYESIAN AND MAXIMUM ENTROPY APPROACH

However, an easy explanation for the origin and derivation of logistic regression exists if one approaches the problem from a Bayesian and Maximum Entropy perspective. Look at the problem in two stages. First there is a symbol manipulation stage to ensure logical consistency. This is then followed by an assignment of numerical values to the symbolic probabilities according to the Maximum Entropy Principle (MEP). The MEP assumes the existence of some model that inserts desired information in order to arrive at the numerical values.

In the first stage, the product and sum rules are called on to manipulate the abstract probability symbols so that the unknown criterion variable of CHD is a function of the patient's age and known lab results. This, of course, results in the simplest form of Bayes's Theorem

$$P(\text{CHD}|\text{AGE,TEST}) = \frac{P(\text{CHD, AGE, TEST})}{P(\text{AGE,TEST})} \tag{3}$$

Since the criterion variable CHD consists of only two values, the denominator in Equation (3) gets expanded by the sum rule into the sum of two joint probabilities,

$$\begin{aligned} &P(\text{CHD=PRESENT}|\text{AGE, TEST}) = \\ &\frac{P(\text{CHD=PRESENT, AGE, TEST})}{P(\text{CHD=PRESENT, AGE, TEST}) + P(\text{CHD=ABSENT, AGE, TEST})} \end{aligned} \tag{4}$$

With this result we have reached the end of the symbol manipulation stage. We choose to leave the probabilities on the right hand side of Equation (4) as joint probabilities rather than transforming them to the typical *likelihood times prior* format.

The rules of probability theory have done their job and we know at the symbolic level that the answer given in Equation (4) must be correct. However, Bayes's Theorem could care less what numerical values are inserted for the joint probabilities on the right hand side as long as they are legitimate probabilities. It will handle all such legitimate probabilities with equal aplomb. Data analysis can proceed no further unless there is some method for assigning numerical values to the joint probabilities. This is where MEP enters the picture.

As explained by Jaynes [4], the MEP is a disciplined method for inserting information into a probability distribution so only that information explicitly mentioned is included and all other unwanted information is excluded. Numerical values are assigned by a model employing constraint functions and the averages of those functions. Many models ranging from the simple to the complex may be proposed as tentative working hypotheses for this numerical assignment. There can be no one true probability assignment, but only assignments conditioned on the information inserted by the tentatively entertained models. The worth of these models is ultimately judged by how closely they predict the actual data that were obtained.

Once we have an MEP model that assigns numerical values, (call it model $\mathscr{M}_k$ and then insert it as supposedly true to the right of the conditioned upon symbol), we can write Bayes's Theorem for the logistic regression set up in Equation (4) as,

$$P(\text{CHD}|\text{AGE, TEST}, \mathscr{M}_k) = \frac{Q_i}{Q_i + Q_j} \tag{5}$$

where Q_i and Q_j are a convenient short notation for the joint probabilities appearing on the right hand side of Bayes's Theorem. This notation indicates that Q_i and Q_j have definite numerical values arising from the MEP.

The MEP provides a formula for computing the numerical values for Q_i and Q_j as,

$$Q_i = \frac{\exp\,[\,\lambda_1 h_1(i) + \lambda_2 h_2(i) + \cdots \lambda_m h_m(i)\,]}{Z(\lambda_1, \lambda_2, \cdots \lambda_m)} \tag{6}$$

and

$$Q_j = \frac{\exp\,[\,\lambda_1 h_1(j) + \lambda_2 h_2(j) + \cdots \lambda_m h_m(j)\,]}{Z(\lambda_1, \lambda_2, \cdots \lambda_m)} \tag{7}$$

Bayes's Theorem can be conveniently rearranged in preparation for the logistic regression by dividing the numerator and denominator of Equation (5) by Q_i.

$$P(\text{CHD}|\text{AGE, TEST}, \mathscr{M}_k) = \frac{1}{1 + \frac{Q_j}{Q_i}} \tag{8}$$

The division of Q_j by Q_i simplifies to

$$\frac{Q_j}{Q_i} = \exp(-Y) \tag{9}$$

where Y is

$$Y = \sum_{l=1}^{m} \lambda_l\,[h_l(i) - h_l(j)] \tag{10}$$

The regression parameters are identified with the Lagrange multipliers and the coding variables with the difference in constraint functions. Thus, in the final step

$$P(\text{CHD}|\text{AGE, TEST}, \mathscr{M}_k) = \frac{1}{1 + \exp(-Y)} \tag{11}$$

and the equivalence with the logistic regression equation is proved.

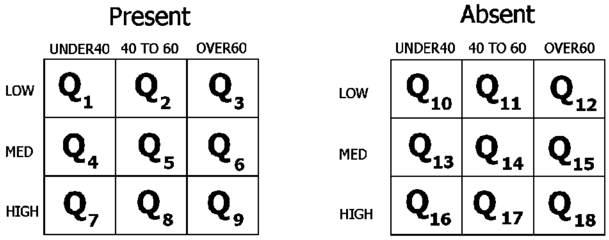
A straightforward application of the sum and product rules for finding the updated probability of a binary variable conditioned on some number of predictor variables, followed by a numerical assignment according to an MEP model, has led to an easy derivation and explanation of the logistic regression equation. Since these are the underlying fundamental principles we would apply for *any* inferencing problem, we see that logistic regression just happens to be a specific example of a general data analysis procedure.

Only the Y term requires some further discussion because of its involvement with the Lagrange multipliers λ_l and constraint functions $h_l(i)$ and $h_l(j)$. These can be matched directly to the regression parameters and coding variables in the standard treatment of logistic regression. A numerical example is presented that shows this correspondence in detail. We then update the probabilities for all the MEP models that were proposed by finding out how well the various models fitted the observed data. The final step is to make a prediction about the probability of coronary heart disease for some new patient based on the patient's known age, lab results, and the data from some previous sample of patients.

THE CONTINGENCY TABLE AND THE CONSTRAINT FUNCTIONS

To further explain the role of the constraint functions and the constraint averages, construct a $2 \times 3 \times 3$ contingency table for the example of calculating the probability of coronary heart disease from two predictor variables. There are two levels for CHD, PRESENT or ABSENT, three levels for the first predictor variable of AGE, UNDER 40, 40 TO 60, or OVER 60, and three levels for the second predictor variable of TEST, and these are LOW, MEDIUM, or HIGH. There are thus a total of $n = 18$ cells in the contingency table. A Q_i value is placed into each one of these cells to indicate the numerical value of the joint probability indexed by that cell. For example, Q_4 is the numerical value as assigned by some MEP model $\mathscr{M}_k$ for the joint occurrence of CHD=PRESENT, AGE=UNDER 40, and TEST=MEDIUM. See Figure 1 for a sketch of the contingency table with the appropriate Q_i filled in for each cell.

Tables 1 and 2 show the values of 17 constraint functions, $h_1(i)$ through $h_{17}(i)$, assigned to each of the 18 cells of the contingency table. In the bottom row, an average



Present

	UNDER40	40 TO 60	OVER60
LOW	Q_1	Q_2	Q_3
MED	Q_4	Q_5	Q_6
HIGH	Q_7	Q_8	Q_9

Absent

	UNDER40	40 TO 60	OVER60
LOW	Q_{10}	Q_{11}	Q_{12}
MED	Q_{13}	Q_{14}	Q_{15}
HIGH	Q_{16}	Q_{17}	Q_{18}

FIGURE 1. A $2 \times 3 \times 3$ contingency table for the logistic regression example. The Q_i notation stands for numerical values assigned to the joint probabilities by a MEP model.

value of each constraint function is given as might be specified by some model. Together with the universal constraint, these 17 constraint functions are a just-determined decomposition of the joint probability over the space of the 18 cells. The universal constraint could be labeled as $h_0(i) = 1$ and $\langle h_0 \rangle = 1$ to enforce the constraint that all probability distributions must satisfy, namely, $\sum_{i=1}^{n} h_0(i)\, Q_i = \langle h_0 \rangle = \sum_{i=1}^{n} Q_i = 1$. The universal constraint does not have to be explicitly listed because the MEP formalism automatically enforces it.

TABLE 1. The first nine of the seventeen constraint functions for the 2 × 3 × 3 contingency table. These are the marginal probability constraints for A, B and C together with the constraints for the AB and AC interactions.

Cell	A	B_1	B_2	C_1	C_2	AB_1	AB_2	AC_1	AC_2
i	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9
1	1	1	0	1	0	1	0	1	0
2	1	0	1	1	0	0	1	1	0
3	1	0	0	1	0	0	0	1	0
4	1	1	0	0	1	1	0	0	1
5	1	0	1	0	1	0	1	0	1
6	1	0	0	0	1	0	0	0	1
7	1	1	0	0	0	1	0	0	0
8	1	0	1	0	0	0	1	0	0
9	1	0	0	0	0	0	0	0	0
10	0	1	0	1	0	0	0	0	0
11	0	0	1	1	0	0	0	0	0
12	0	0	0	1	0	0	0	0	0
13	0	1	0	0	1	0	0	0	0
14	0	0	1	0	1	0	0	0	0
15	0	0	0	0	1	0	0	0	0
16	0	1	0	0	0	0	0	0	0
17	0	0	1	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0
$\langle h_l \rangle$	.20	.20	.50	.33	.33	.03	.09	.05	.06

So in specifying an MEP model $\mathscr{M}_k$ that assigns numerical values to the 18 cells of the contingency table, the m appearing in Equation (10) could range from $m = 0$ to $m = n - 1$. Whenever $m < n - 1$, there is an underdetermined problem and the MEP serves the very important role of resolving any remaining ambiguity about the assignment. The MEP inserts the information from m constraint functions and constraint averages specified by a model into a joint probability distribution, *and inserts no other information other than that explicitly identified*. This is the insurance policy that we take out when we rely on the MEP to assign numerical values.

For ease of presentation, let A stand for the binary criterion variable of CHD, let B stand for the first predictor variable of AGE, and let C stand for the second predictor variable of TEST. The constraint functions $h_1(i)$ through $h_{17}(i)$ code for the marginal probabilities of the A, B, and C variables, as well as for all possible double interactions, AB, AC, and BC, which are also finer grained marginal probabilities, and finally for the one triple interaction, ABC. Together, these marginal probabilities and single cell probabilities constitute the $m = 17$ constraint functions.

TABLE 2. The last eight of the seventeen constraints for the 2 × 3 × 3 contingency table. These are the *BC* and *ABC* interactions.

Cell	B_1C_1	B_1C_2	B_2C_1	B_2C_2	AB_1C_1	AB_1C_2	AB_2C_1	AB_2C_2
i	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}	h_{16}	h_{17}
1	1	0	0	0	1	0	0	0
2	0	0	1	0	0	0	1	0
3	0	0	0	0	0	0	0	0
4	0	1	0	0	0	1	0	0
5	0	0	0	1	0	0	0	1
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0
11	0	0	1	0	0	0	0	0
12	0	0	0	0	0	0	0	0
13	0	1	0	0	0	0	0	0
14	0	0	0	1	0	0	0	0
15	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0
$\langle h_l \rangle$	.068	.067	.167	.167	.008	.009	.022	.027

As an example, look at the second column showing the constraint function $h_1(i)$ for $i = 1$ to 18. This codes for the marginal probability of A, or CHD. Thus, $h_1(i) = 1$ for $i = 1$ to 9 and $h_1(i) = 0$ for $i = 10$ to 18. The average of this first constraint function is specified by a model as $\langle h_1 \rangle = .20$.

Therefore,

$$\begin{aligned} \sum_{i=1}^{18} h_1(i)\, Q_i &= \langle h_1 \rangle \\ Q_1 + Q_2 \cdots + Q_9 &= .20 \end{aligned}$$

and the marginal probability for A is inserted by this model as the one piece of information (along with, of course, the universal constraint) into a joint distribution. Refer back to Figure 1 to see that $\sum_{i=1}^{9} Q_i$ is the marginal probability for the first level of variable A, the presence of coronary heart disease. The second level of A for this particular model is automatically determined by the sum rule as .80.

A similar explanation holds for the constraint functions $h_2(i)$ through $h_5(i)$. These code for the marginal probabilities of B and C, or AGE and TEST. The only difference is that since AGE and TEST exist at three levels, two constraint functions are required for each of these predictor variables. By constructing the constraint functions $h_1(i)$ through $h_5(i)$ and their associated averages, we have generated models that insert information into a joint probability distribution about the marginal probabilities for A, B and C, or CHD, AGE and TEST.

However, it turns out that all such models that use $h_1(i)$ through $h_5(i)$ exhibit independence between the criterion variable A and the predictor variables B and C. We have not built any constraint functions that allow for a relationship between A and B or A and C, so that the probability for A can change when conditioned on B and/or C. By constructing constraint functions $h_6(i)$ and $h_7(i)$, we specify models that allow for such an association between A and B, and functions $h_8(i)$ and $h_9(i)$ do the same for A and C. The constraint functions $h_{10}(i)$ through $h_{13}(i)$ permit models to be considered that also express an association between B and C. This takes care of all the double interactions. Finally, the last four constraint functions and their averages, $h_{14}(i)$ through $h_{17}(i)$, capture a relationship among all three variables at once, or what we call the triple interaction ABC.

The interaction constraints are found by multiplying the separate constraints that make up the interaction. For example, the AB_1 interaction, $h_6(i)$, is found from the separate A and B constraint functions as in $h_6(i) = h_1(i) \times h_2(i)$. Referring once again to the actual values of $h_6(i)$ in Table 1, we see that

$$\begin{aligned}\sum_{i=1}^{18} h_6(i)\, Q_i &= \langle h_6 \rangle \\ Q_1 + Q_4 + Q_7 &= .03\end{aligned}$$

This is the marginal probability for CHD=PRESENT and AGE=UNDER 40 over all three levels of TEST.

NUMERICAL EXAMPLE

Suppose we wanted to find the revised probability for the presence of coronary heart disease given knowledge of a C-reactive protein test and the age of the patient. The particular patient we are interested in obtained a MEDIUM score on the diagnostic test and is under 40 years of age. By specifying some MEP model, we can assign numerical values to all 18 Q_i values. The symbolic manipulation rules then tell us that

$$\begin{aligned}P(\text{CHD=PRESENT}\,|\,\text{AGE=UNDER 40, TEST=MEDIUM}, \mathscr{M}_k) &= \frac{Q_4}{Q_4 + Q_{13}} \\ &= \frac{1}{1 + Q_{13}/Q_4}\end{aligned}$$

For the sake of the numerical example to follow, suppose that the model $\mathscr{M}_k$ consists of $h_1(i)$ through $h_9(i)$ and their associated constraint averages. That is, the model inserts the information about the marginal probabilities for A, B, and C and also inserts the information about the two double interactions, AB and AC. This is equivalent to tentatively entertaining a model where both B and C influence the probability of A.

In our example, we are interested in a model where both the test results and the age of the patient have an impact on the probability of coronary heart disease. In working out the details of this example according to the Bayesian and Maximum Entropy principles, we will discover that the results of the conventional approach are exactly the same, but without providing the all important rationale to support and justify the data analytic procedures.

From Equation (10), and the fact that we are using a model where $m = 9$, we know that

$$\begin{aligned} Y &= \sum_{l=1}^{9} \lambda_l \left[h_l(4) - h_l(13)\right] \\ &= \lambda_1 \left[h_1(4) - h_1(13)\right] + \lambda_2 \left[h_2(4) - h_2(13)\right] + \cdots + \lambda_9 \left[h_9(4) - h_9(13)\right] \\ &= \lambda_1 \left[1 - 0\right] + \lambda_2 \left[1 - 1\right] + \cdots + \lambda_9 \left[1 - 0\right] \\ &= \lambda_1 + \lambda_6 + \lambda_9 \end{aligned}$$

The functional values of each $h_l(i)$ were pulled from Table 1. The values of the Lagrange multipliers are found by a numerical routine on the computer [5] to be

$$\begin{aligned} \lambda_1 &= -.620 \\ \lambda_6 &= -.723 \\ \lambda_9 &= -.522 \end{aligned}$$

From Equation (9),

$$\begin{aligned} \frac{Q_{13}}{Q_4} &= \exp\left(-Y\right) \\ &= \exp\left(1.865\right) \end{aligned}$$

Therefore, the updated probability of coronary heart disease for this patient is

$$\begin{aligned} &P(\text{CHD=PRESENT}|\text{AGE=UNDER 40, TEST=MEDIUM}, \mathscr{M}_k) \\ &\qquad = \frac{1}{1 + \exp\left(1.865\right)} \\ &\qquad = .1341 \end{aligned}$$

Now, how would this analysis be carried out conventionally? The commercially available computer programs for logistic regression tell you to code the discrete predictor variables as "dummy" indicator variables as shown below in Table 3. The regression equation for two predictor variables then becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

The maximum likelihood estimates of the parameters are returned by the program as

$$\begin{aligned}
\beta_0 &= -.620 \\
\beta_1 &= -.723 \\
\beta_2 &= -.505 \\
\beta_3 &= -.740 \\
\beta_4 &= -.522
\end{aligned}$$

For the particular settings of the indicator variables of UNDER 40 and MEDIUM

$$\begin{aligned}
Y &= -.620 + (-.723 \times 1) + (-.505 \times 0) + (-.740 \times 0) + (-.522 \times 1) \\
&= -.620 - .723 - .522 \\
&= -1.865
\end{aligned}$$

At this point, it is obvious that the answer obtained from a program implementing the conventional approach is exactly the same as just derived in detail by the combined Bayesian and MEP approach.

TABLE 3. The "dummy" coding variables used for the two predictor variables in a conventional logistic regression program.

AGE		
X_1	X_2	**Levels**
1	0	UNDER 40
0	1	40 TO 60
0	0	OVER 60

TEST		
X_3	X_4	**Levels**
1	0	LOW
0	1	MEDIUM
0	0	HIGH

THE DATA AND REVISION OF MODEL PROBABILITIES

No mention has been made so far of any actual data. This is because they have not yet been needed. The probability of CHD can be calculated quite readily given some feasible model for the numerical assignment of the joint probabilities. We have seen these calculations in the previous sections. Those who experience some discomfort at seeing the data appear so late on the scene, harbor, I suspect, a lingering attachment to probability defined as a frequency. When the data are not around to anchor their true feelings about a probability, they start to become uncomfortable.

The only role of the data (admittedly though, a very important role) is to update the probabilities for all the models under consideration. Figure 2 shows some frequency counts placed into the cells of the contingency table from a hypothetical experiment.

1,000 patients were examined for the presence of CHD and their age and results of the diagnostic test were recorded as well. Each patient was classified into one, and only one, of the 18 cells of the table. Various marginal totals are also shown. These data will be given the notation of $\mathscr{D}$ with the acknowledgment that there is no uncertainty attached to the counts in the cells when these have been accurately recorded.

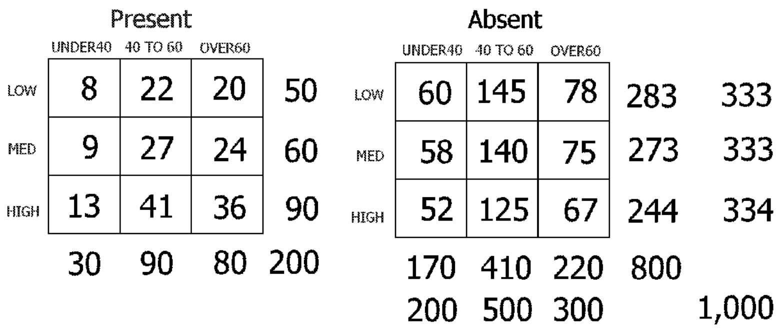
We can now explain the somewhat mysterious origin of the constraint averages, $\langle h_l \rangle$, that appear in the models. As an example, the average for the first constraint was set by some model as $\langle h_1 \rangle = .20$ as shown in the bottom row of Table 1. The marginal frequency count for the presence of CHD was 200, so this particular constraint average selected by the model uses the relative frequency of 200/1000=.20.

Other models might very well choose to set $\langle h_1 \rangle = .19$ or $\langle h_1 \rangle = .21$, or any other feasible value for that matter. But if we set $\langle h_1 \rangle$ to the marginal frequency of A as revealed by the actual empirical data, we are doing within the MEP the same thing as what the conventional approach does when it picks a maximum likelihood estimate. Jaynes demonstrates this interesting fact in [6]. This also clears up the coincidence that the Lagrange multipliers and the maximum likelihood estimates of the regression parameters had the same values, as was discovered in the numerical example.

Blower [7], following the lead of Jaynes, proved that the likelihood ratio could be worked into

$$\frac{P(\mathscr{D}|\mathscr{M}_k)}{P(\mathscr{D}|\mathscr{M}_F)} = \exp\,(N\Delta H) \tag{12}$$

where ΔH is the difference in information entropy between any two models. For example, there is a model, call it model $\mathscr{M}_F$, that assigns probabilities exactly equal to the normed frequencies, $f_i = N_i/N$. Model $\mathscr{M}_F$ is Jaynes's "sure-thing" hypothesis. This



	Present					Absent				
	UNDER40	40 TO 60	OVER60			UNDER40	40 TO 60	OVER60		
LOW	8	22	20	50	LOW	60	145	78	283	333
MED	9	27	24	60	MED	58	140	75	273	333
HIGH	13	41	36	90	HIGH	52	125	67	244	334
	30	90	80	200		170	410	220	800	
						200	500	300		1,000

FIGURE 2. Observed frequency counts from a hypothetical experiment involving $N = 1,000$ patients. Each patient was placed into one and only one of the 18 cells of the contingency table depending upon the joint occurrence of CHD, AGE, and TEST.

difference in information entropies turns out to equal

$$\Delta H = \sum_{i=1}^{n} f_i \ln\left(\frac{f_i}{Q_i}\right) + \sum_{l=1}^{m} \lambda_l \left(\overline{h}_l - \langle h_l \rangle\right) \tag{13}$$

By setting the constraint averages $\langle h_l \rangle$ equal to the sample averages $\overline{h}_l$ in Equation (13), the second term is eliminated and the difference in entropies is a cross-entropy involving the frequencies from $\mathscr{M}_F$ and the Q_i values from $\mathscr{M}_k$.

Equally important is our ultimate goal of *predicting* CHD for a *new* patient not in the database of 1,000 previously studied patients. Using the same symbolic manipulation rules as before, we find that the posterior predictive probability that captures an updated state of knowledge about a new patient's chance of CHD given that we know the age, the test results for the new patient, and the previous data sample, is

$$\begin{aligned} &P(\mathrm{CHD}_{New}|\mathrm{AGE}_{New}, \mathrm{TEST}_{New}, \mathscr{D}) = \\ &\sum_{k=1}^{M} P(\mathrm{CHD}_{New}|\mathrm{AGE}_{New}, \mathrm{TEST}_{New}, \mathscr{M}_k)\, P(\mathscr{M}_k|\mathscr{D}) \end{aligned} \tag{14}$$

In order to find the posterior predictive probability, we average each model's individual prediction weighted by the posterior probability of the model in question. The first term in the sum is the Bayes's theorem calculation based on the numerical assignment to the joint probabilities by the MEP model, $\mathscr{M}_k$, just as we have been discussing in the previous section. It does not depend on the past data, but depends only on the specified model. In essence, assumption of the truth of a particular model supersedes any amount of prior data. What does depend on the past data is the posterior probability for each model being considered.

The ratio of the posterior probability of any model under consideration in the Bayesian model averaging sum of Equation (14), (generically model $\mathscr{M}_1$), to the posterior probability of any other model $\mathscr{M}_2$ is

$$\frac{P(\mathscr{M}_1|\mathscr{D})}{P(\mathscr{M}_2|\mathscr{D})} = \frac{P(\mathscr{D}|\mathscr{M}_1)\, P(\mathscr{M}_1)}{P(\mathscr{D}|\mathscr{M}_2)\, P(\mathscr{M}_2)} \tag{15}$$

All models will be on an equal footing before the data have been collected so that

$$\frac{P(\mathscr{M}_1)}{P(\mathscr{M}_2)} = 1 \tag{16}$$

and the ratio of posterior probabilities reduces to a likelihood ratio. However, we will introduce a term (the Bayesian Information Criterion) to penalize models with more parameters [8].

Here is an example of the computation to find the relative weight for any two models. Take one of the models $\mathscr{M}_1$ on the right hand side of Equation (14) to be the same model as in the previous example. This model, recall, included information about the marginal probabilities of A, B and C plus the interaction terms AB and AC. The wish is to tentatively entertain the hypothesis that there was some association between the criterion variable A and both of the predictor variables B and C. The second model will claim, instead, that there is an association with only the predictor variable B. This first model has $m = 9$ constraint functions and associated constraint averages. For notational purposes, label the model as $\mathscr{M}_1 \equiv \mathscr{M}_{AB+AC}$.

Compare this model to another model $\mathscr{M}_2$ which we will take to be a model that inserts information about an association between just A and B, together with the information about the marginal probabilities. Thus, this model incorporates only the AB interaction, that is, the first seven constraint functions are used and $m = 7$. For notational purposes, label this model as $\mathscr{M}_2 \equiv \mathscr{M}_{AB}$.

As explained above, the constraint averages in both of these models were set to the sample averages so that the results would correspond to the conventional maximum likelihood treatment. ΔH loses the second term in Equation (13) above and we can now proceed to the computation. This difference in entropies can be expressed as a cross-entropy expression where the cross-entropy reflects the closeness of the frequency data to each model's Q_i values.

$$\Delta H = \sum_{i=1}^{n} f_i \ln \left(\frac{Q_i^{AB+AC}}{Q_i^{AB}} \right) \tag{17}$$

The ratio of the posterior probability of our model $\mathscr{M}_{AB+AC}$ to the less complex model $\mathscr{M}_{AB}$, and with equal prior probabilities for both models, becomes the likelihood ratio.

$$\begin{aligned}
\frac{P(\mathscr{M}_{AB+AC}|\mathscr{D})}{P(\mathscr{M}_{AB}|\mathscr{D})} &= \frac{P(\mathscr{D}|\mathscr{M}_{AB+AC})}{P(\mathscr{D}|\mathscr{M}_{AB})} \\
\frac{P(\mathscr{D}|\mathscr{M}_{AB+AC})}{P(\mathscr{D}|\mathscr{M}_{AB})} &= \exp\left(N\Delta H\right) \\
\Delta H &= \sum_{i=1}^{18} \frac{N_i}{N} \ln \left(\frac{Q_i^{AB+AC}}{Q_i^{AB}} \right) \\
\Delta H &= \frac{8}{1000} \ln \left(\frac{.0075}{.0100} \right) + \frac{22}{1000} \ln \left(\frac{.0225}{.0300} \right) + \cdots \\
& \quad + \frac{67}{1000} \ln \left(\frac{.066917}{.073333} \right) \\
&= .00786 \\
N\Delta H &= 7.858 \\
\text{BIC} &= 5.070 \\
\frac{P(\mathscr{M}_{AB+AC}|\mathscr{D})}{P(\mathscr{M}_{AB}|\mathscr{D})} &= \exp\left(7.858 - 5.070\right) \\
&= \approx 16
\end{aligned}$$

The model $\mathscr{M}_{AB+AC}$ fits the actual data very well as can be seen by comparing the values of f_i and the numerator Q_i values in the sum comprising ΔH. Therefore, $\mathscr{M}_{AB+AC}$'s prediction will receive a relative weighting in Equation (14) about 16 times that of model $\mathscr{M}_{AB}$'s prediction which fit the data rather less well. This calculation of the relative weighting includes a BIC penalty term for the two extra parameters in model $\mathscr{M}_{AB+AC}$. In like manner, the model space would be explored by a computer program with the predictions made by all the models visited weighted by each model's relative posterior probability. The final average would be a good estimate of the posterior predictive probability for the presence of CHD given knowledge of the age of the patient, his or her test results, and the outcomes of the 1,000 previous patients.

SUMMARY

Jaynes [9] has presented the theoretical foundation for scientific inference as a generalization of ordinary logic. This paper is a practical application of his approach to the well-known data analytic technique of logistic regression. Because the formal manipulation of probabilities is simple and direct, we have concentrated on the numerical assignment to joint probabilities via models specified within the MEP formalism. By spelling out in detail the nature of the constraint functions and constraint averages for independence models and, in addition, models that postulated some relationship between the criterion and predictor variables, we could discuss the origin and rationale for logistic regression. As an interesting side benefit, the maximum likelihood solution of the conventional approach was seen to be numerically equal to the MEP solution that used the data as constraint averages. By employing Jaynes's advice, there is no reason why other data analysis procedures will not also yield to an easy rationale for what were previously rather arcane justifications, or failing that, they will be judged to be essentially deficient.

ACKNOWLEDGMENTS

I would like to extend my appreciation to Dr. Ken Ford, the Director of the Institute for Human and Machine Cognition, for supporting my efforts to develop a general theory of inference for scientific data analysis. Much of this work was carried out while on assignment to the Naval Aerospace Medical Research Laboratory and I thank the Commanding Officer and staff for providing a pleasant environment in which to conduct this research.

REFERENCES

1. Hosmer, D.W. and Lemeshow, S., *Applied Logistic Regression*, 2nd Edition, John Wiley & Sons, New York, 2000.
2. Bishop, C.M., *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, UK, 1995, pp. 82–83.
3. Dennison, D. G. T., Holmes, C. C., Mallick, B. K., Smith, A. F. M. *Bayesian Methods for Nonlinear Classification and Regression*, John Wiley & Sons Ltd., Chichester, UK, 2002.
4. Jaynes, E.T., "Brandeis Lectures," Chapter 4 in *Papers on Probability, Statistics and Statistical Physics*, edited by R. D. Rosenkrantz, Kluwer Academic Publishers Norwell, MA, 1989.
5. Fletcher, R., *Practical Methods of Optimization*, John Wiley & Sons Ltd., Chichester, UK, 1987, pp. 222–223.
6. Jaynes, E.T., "Where Do We Stand on Maximum Entropy?", Chapter 10 in *Papers on Probability, Statistics and Statistical Physics*, edited by R. D. Rosenkrantz, Kluwer Academic Publishers, Norwell, MA, 1989, pp. 270–271.
7. Blower, D. J., "A Bayesian Approach to Navy Enlisted Compensation Models," *Final Report to Naval Personnel Research Studies and Technology Command*, 2001.
8. Leonard, T. and Hsu, J., *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge University Press, Cambridge, UK, 1999, pg. 8.
9. Jaynes, E.T., *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.