MAXIMUM ENTROPY FOR HYPOTHESIS FORMULATION, ESPECIALLY FOR MULTIDIMENSIONAL CONTINGENCY TABLES

By I. J. Good

Institute for Defense Analyses and Admiralty Research Laboratory

0. Summary. The principle of maximum entropy, together with some generalizations, is interpreted as a heuristic principle for the generation of null hypotheses. The main application is to m-dimensional population contingency tables, with the marginal totals given down to dimension $m - r$ ("restraints of the rth order"). The principle then leads to the null hypothesis of no "rth-order interaction." Significance tests are given for testing the hypothesis of no rth-order or higher-order interaction within the wider hypothesis of no sth-order or higher-order interaction, some cases of which have been treated by Bartlett and by Roy and Kastenbaum. It is shown that, if a complete set of rth-order restraints are given, then the hypothesis of the vanishing of all rth-order and higher-order interactions leads to a unique set of cell probabilities, if the restraints are consistent, but not only just consistent. This confirms and generalizes a recent conjecture due to Darroch. A kind of duality between maximum entropy and maximum likelihood is proved. Some relationships between maximum entropy, interactions, and Markov chains are proved.

1. Introduction. This paper deals with a certain heuristic method of generating null hypotheses. By a "null hypothesis" we mean a hypothesis that is liable to be worth testing. The method might also be of value for curve-fitting and surface-fitting. Most of the paper however is concerned with interactions in multidimensional contingency tables and in Markov chains.

Let X be a random variable whose (physical) probability distribution is not completely given. Of all the available distributions there will usually be one of maximum entropy, i.e., of maximum uncertainty. We shall consider some implications of the following principles.

Principle of Maximum Entropy. *Let X be a random variable whose distribution is subject to some set of restraints. Then entertain the null hypothesis that the distribution is the one of maximum entropy, subject to these restraints.*

Historical note. Jaynes [23] used precisely the same formal procedure, but with a different interpretation, since he was concerned with the choice of distributions of nonphysical probabilities, sometimes called "credibilities", as a method of replacing the use of a Bayes-Laplace postulate of a uniform distribution of credibility. He applied the method to statistical mechanics following earlier writers such as Boltzmann [2] and Gibbs. For the application to statistical

Received June 12, 1962; revised March 28, 1963.

mechanics, we suspect that our interpretation of the principle of maximum entropy would have been logically more appropriate. That is, the principle of maximum entropy generates much of statistical mechanics as a *null hypothesis*, to be tested by experiment.

Prima facie reasons for entertaining the principle of maximum entropy, and some generalizations. The mere fact that the principle of maximum entropy generates classical statistical mechanics, as a null hypothesis, would be a sufficient reason for examining many of its implications in mathematical statistics.

The principle can be expressed in slightly different words, such as: *entertain the hypothesis, H, that maximizes the expected amount of selective information per observation.* Or again: *entertain the hypothesis, H, for which the expected weight of evidence per observation, as compared with the hypothesis of a uniform distribution, is a minimum*, i.e., in the discrete case, $\sum p_i \log(np_i)$ is a minimum, where $p_1, p_2, \cdots, p_n$ are the probabilities of the n mutually exclusive and exhaustive possible values of the random variable. (See below for a definition of "weight of evidence.")

The various descriptions of H are simple enough to justify the selection of H as a null hypothesis, in accordance with Occam's razor. It is true that Occam's razor is usually invoked in order to justify the selection of a scientific theory, but the logical status of a null hypothesis is much the same as that of a scientific theory in the sense that both are conjectured to be approximately valid, and are considered to be worth testing.

It could be objected that, especially for a continuous distribution, entropy is somewhat arbitrary, since it is variant under a transformation of the independent variable. (In Shannon's theory of information, this lack of invariance is immaterial since it cancels out when one calculates the expected amount of information provided by one distribution concerning another one, if the first distribution is absolutely continuous with respect to the second one.) To this objection we have a few replies:

(a) The principle of maximum entropy is intended only as a heuristic principle for generating null hypotheses for consideration. Statistical theory is poor in such suggestions: hypotheses are usually assumed to be formulated before statistical theory is invoked. This is a weakness in statistical theory, regarded as a part of scientific method, consequently some new results in this direction should be of interest.

(b) The independent variable should be taken in a natural or simple way, or at least the most natural and simplest ways should be tried first. This again is an application of Occam's razor, and it is not peculiar to the principle of maximum entropy that naturalness and simplicity have not been adequately formalized.

(c) The entropy of a continuous distribution can be thought of as that of a discrete distribution, if the independent variable is measured to only limited accuracy. (For example [20], p. 114.) The accuracy that is attainable, in any physical problem, might vary from one part of the range to another: *it certainly would vary if the range were infinite.* One natural independent variable is the one

for which the range is divided up into the smallest number of intervals such that two adjacent intervals are only just discernible by the apparatus at your disposal. Then a physically natural independent variable will be one in which each of these intervals is of equal length. The principle of maximum entropy would then lead to the null hypothesis in which each just discernible interval is assigned equal probability. (Compare Lindley [32] and Good [18].)

(d) More generally, if, not allowing for the restraints, we have some reason for regarding a distribution, of density function $g(\cdot)$, as *ausgezeichnet*, we could replace the entropy by Turing's expression,

$$\int f(x) \log \frac{f(x)}{g(x)}\, dx = \mathcal{E}\left\{\log \frac{f(x)}{g(x)} \,\middle|\, f(\cdot)\right\}$$

which has been called an "expected log-factor" (logarithm of a Bayes factor) or *expected weight of evidence* in favour of $f(\cdot)$ as against $g(\cdot)$ per observation, given $f(\cdot)$; or the *cross-entropy* between f and g. ([11], pp. 72 and 75, [9], [14], [26].) We should then have the following generalization of the principle of maximum entropy:

PRINCIPLE OF MINIMAL DISCRIMINABILITY. *Let X be a random variable whose distribution is subject to some set of restraints. Suppose that, before the restraints were known, there was some distribution that seemed reasonable to entertain as a null hypothesis, called an initially ausgezeichnet hypothesis. This hypothesis is perhaps refuted by the constraints. Then, in view of the restraints, entertain the null hypothesis that, if true, can be discriminated from the ausgezeichnet hypothesis at the minimum rate, i.e., for which the expected weight of evidence per observation is least.*

If the ausgezeichnet distribution is not refuted by the constraints, then of course the null hypothesis picked out by the above principle will be the *ausgezeichnet* distribution itself. (See example (i) below.)

B. O. Koopman has mentioned (private communication) that cross-entropy can be used in the foundations of statistical mechanics for non-equilibrium conditions, so that we have another reason for regarding the principle of minimal discriminability as a natural generalization of that of maximum entropy. ([9], [14].)

When there is no distribution that is ausgezeichnet, then the question will arise whether the independent variable, x, has been chosen in the most natural manner. If it has, then the uniform distribution and the principle of maximum entropy will be reasonable. If this seems arbitrary, it should be remembered that this degree of arbitrariness is present whenever entropy is discussed, not merely in the present application.

For the sake of simplicity, this generalized form of the principle of maximum entropy will not be used in the present paper.

A different kind of generalization can be formulated for the purpose of generating composite hypotheses:

PRINCIPLE OF MAXIMUM CONDITIONAL ENTROPY. *Suppose that for some simply defined class of restraints, the principle of maximum entropy leads to a simply*

expressible set of null hypotheses. Then this whole class can reasonably be set up as a composite null hypothesis, even when the restraints are not operative.

Note that this second principle involves an additional element of judgment in its application, since it refers again to the notion of "simplicity" for which there still exists no fully satisfactory formal definition.

A principle of *minimal conditional discriminability* could now be stated, by combining the two generalizations. Its statement is obvious and will be omitted.

Our main application of the principle of maximum entropy is to m-way ($= m$-dimensional) contingency tables. We are led to consider null hypotheses of no "rth-order interaction". It seems to be useful in the analysis to make use of the m-dimensional discrete Fourier transform of the logarithms of the cell probabilities.

Section 2. concludes with some significance tests for testing the hypothesis of no rth-order or higher-order interaction within the wider hypothesis of no sth-order or higher-order interaction, some cases of which had previously been treated by Bartlett [1] and by Roy and Kastenbaum [37].

Section 3. discusses the relationship of the work with (i) two papers that appeared very recently, Plackett [35] and Darroch [5]: a proof of a conjecture of Darroch's is found to be implicit in the present paper; (ii) latent class analysis; (iii) interactions for Markov chain; (iv) contingency tables for a small sample; (v) another method for generating null hypotheses; (vi) orthogonal interactions.

A potential application of considerable interest, for multidimensional contingency tables, is to problems of pattern recognition, such as to the recognition of phonemes in speech or of the letters of the alphabet from script. If each letter is classified by m attributes, taking respectively $d_1, d_2, \cdots, d_m$ values, then a sample could be summarised by means of a $d_1 \times d_2 \times \cdots \times d_m \times 26$ contingency table. If the sample size is not very large, nearly all the cell entries in the table might be very small, so that a method is required for estimating the population probabilities. In principle this would lead to a method for determining the conditional probabilities of the letters given the values of the m attributes. It is hoped that the present paper will contribute to the solution of this problem.

2. Examples.

Example (i). *Finite discrete distribution.* Let $X = i$ with probability p_i ($i = 1, 2, \cdots, n$; $p_1 + p_2 + \cdots + p_n = 1$), where the p_i's are unknown. If we maximize the entropy $-\sum p_i \log p_i$ subject to the restraint $\sum p_i = 1$ we find that $p_1 = p_2 = \cdots = p_n = 1/n$. The null hypothesis thrown up for consideration by the principle of maximum entropy is thus that the probabilities p_i are all equal. *Formally*, this is the same as the usual Bayes-Laplace postulate, but the meaning is logically entirely different here, since our null hypothesis states that the *physical* probabilities are all equal. For a discussion of the distinction between kinds of probability see, for example, Good [16].

If the principle of maximum entropy is generalized to a principle of expected weight of evidence, relative to a distribution (q_i), as described above, then the

null hypothesis that is thrown up is that the p_i's are equal to the q_i's. This indeed is a reasonable selection of a null hypothesis.

Example (ii). *Continuous distribution.* Let X be a random variable with a continuous n-dimensional distribution in unbounded n-space, and let the second moments be assigned. Then Shannon [39], p. 629, mentions that the X of maximum entropy is Gaussian with zero means. This then is the null hypothesis for the distribution of X that is suggested by the principle of maximum entropy. If the covariance matrix is singular, then X is restricted to a manifold of dimensionality less than n, and the entropy must be defined with respect to the reduced space. See also Good and Doog [20], p. 122.

Even if the second moments are not given, the principle of maximum *conditional* entropy suggests as a null hypothesis that the distribution is Gaussian of mean zero. A null hypothesis generated for a time series is that it is Gaussian.

Example (iii). *Distribution with a finite number of moments assigned.* Let us again consider a discrete distribution, (p_i), where i can take n distinct values; and suppose that we have assigned values for the moments

$$\sum_i i^r p_i = \mu_r' \quad (r = 0, 1, \cdots, k; \qquad k < n).$$

Assume further that these restraints are consistent and do not force any of the p_i's to vanish, i.e., the restraints are not *only just consistent.* In other words, the set of consistent restraints contains an open neighbourhood of the point $(\mu_0', \mu_1', \cdots, \mu_k')$. We wish to maximize $y = -\sum p_i \log p_i$ subject to the restraints. The Lagrange expression is

$$-\sum p_i \log p_i + \lambda_0 + \lambda_1 \sum i p_i + \cdots + \lambda_k \sum i^k p_i$$

where λ_r $(r = 0, 1, \cdots, k)$ are undetermined multipliers. The equations for a stationary point are

$$1 + \log p_i = \lambda_0 + i\lambda_1 + i^2\lambda_2 + \cdots + i^k\lambda_k.$$

Now the Hessian matrix of second derivatives of y is a diagonal matrix that is negative definite. So there can be at most one stationary point and this must be an absolute maximum if it exists. In order to prove that it does exist, it is enough to show that the maximum value of y cannot occur on the boundary of the region of definition, i.e., at a point where any one of the p_i's vanishes. This can be shown by precisely the same argument as is used in Example (vii) below for proving that y cannot be a maximum on its boundary of definition. (The proof depends on the finiteness of n.) It follows that p_i is of the form of the exponential of a polynomial, $\exp(c_0 + c_1 i + c_2 i^2 + \cdots + c_k i^k)$, where the coefficients $c_0, c_1, \cdots, c_k$ are uniquely defined by equating the moments to the assigned values.

The above argument generalizes at once to the case where i^r is replaced by an arbitrary function, $\varphi_r(i)$, where $\varphi_0(i) = 1$, so that μ_r' becomes a generalized moment. The argument also generalizes to discrete distributions of several variables, $i_1, i_2, \cdots, i_m$.

The result shows as a by-product that if the assigned moments (or generalized moments) are consistent, and not "only just" consistent, then it is *possible* to satisfy the restraints by taking $\log p_i$ as a polynomial of degree k, or as a generalized polynomial, $c_0 + c_1\varphi_1(i) + \cdots + c_k\varphi_k(i)$.

Note that if any of the moments is unspecified, the corresponding term in the polynomial is missing.

Formally, and with considerable risk, we may make the transition to continuous distributions. If the probability density is $f(x_1, x_2, \cdots, x_m) = f(x)$ and if we are given a finite number of moments, then the null hypothesis generated by the principle of maximum entropy is that $f(x)$ is the exponential of a polynomial, in which there is a term in $x^k = x_1^{k_1}x_2^{k_2} \cdots x_m^{k_m}$ (or in $\varphi_k(x)$) if the expectation of x^k (or of $\varphi_k(x)$) is assigned. A necessary condition for the validity of this formalism is that the integral of the exponential of relevant polynomials should converge over the range of values of x. For example, in the one-dimensional case, the moment of highest assigned order must be of even order, otherwise the upper bound of the entropy is not attainable. This difficulty can occur in the discrete case also, when the values of i range from $-\infty$ to $+\infty$.

In problems of curve-fitting, it is not usually reasonable to consider an infinite range for the independent variable. Any extrapolation of a fitted curve outside the range of the actual observations, and especially into the extreme tails, is likely to be very tentative. Consequently it is reasonable to use only a finite range for the independent variable, and this has the advantage of avoiding the above difficulty. But, in view of this difficulty, it seems safer to work with moments of order $1, 2, 3, \cdots, 2k$, where $2k$ is even.

If we restrict our attention to ordinary moments, the implications for curve-fitting are as follows.

Suppose it is considered reasonable to summarise the features of the possible distributions by means of moments up to an even order. Then the probability-density curves thrown up by the principle of maximum entropy are exponentials of even-degree polynomials $\exp(c_0 + c_1x + \cdots + c_{2k}x^{2k})$, where c_0 can be thought of as a normalizing constant. If the data consists of observations $x_1, x_2, \cdots, x_N$, then the log-likelihood of the above density is

$$\lambda_{2k} = \sum_{i=0}^{2k} \sum_{j=1}^{N} c_i x_j^i ,$$

and the coefficients can be determined by maximum likelihood. The hypothesis, H_{2k}, that $2k$ is the highest order term, can be tested within the wider hypothesis H_{2k+2l} by means of the likelihood-ratio criterion $2\hat{\lambda}_{2k+2l} - 2\hat{\lambda}_{2k}$, where the caps indicate that the log-likelihoods have been maximized. This criterion has, for large N, approximately a chi-squared distribution with $2l$ degrees of freedom, when the hypothesis H_{2k} is true.

It should be noted that if we are driven to reject say H_2, H_4, and H_6 this would be evidence against the *entire* sequence $H_2, H_4, H_6, \cdots$. This would be especially true when some rival method of curve-fitting is also considered to be

reasonable. For example, a linear mixture of two normal distributions might, in a given practical situation, seem more reasonable than the exponential of a quartic. Here again the likelihood-ratio test could be used, in the generalized forms suggested by Cox [4]. Both here and elsewhere, the generalized likelihood ratio, generalized in a different manner, could be employed if desired [17].

For surface, or hyper-surface, fitting, one could define H_{2k} as the composite hypothesis whose parameters are *all* the moments through the $2k$th. The criterion $2\hat{\lambda}_{2k+2} - 2\hat{\lambda}_{2k}$, for example, would have $4k + 1$ degrees of freedom in two dimensions, and

$$(d + 2k - 2)!\,(d + 4k - 1)/(2k)!\,(d - 1)!$$

in d dimensions.

Example (iv). *Two-way "population" contingency table, with assigned marginal probabilities.* By a "population" contingency table we mean an array (p_{ij}) of population probabilities, with row sums $(p_{i,})$ and column sums $(p_{,j})$. (An acute accent implies here and later that the corresponding suffix has been summed out.) It is a question now of maximizing

$$-\sum_{ij} p_{ij} \log p_{ij}$$

subject to the restraints

$$\sum_{j} p_{ij} = p_{i,}\,, \qquad \sum_{i} p_{ij} = p_{,j}\,.$$

Note that the principle of maximum entropy can here be expressed by saying that the expected amount of mutual information between rows and columns is minimized. (Cf. Hartmanis, [22], Lewis, [31].)

An application of Lagrange's method of undetermined multipliers (see the next example) leads to the null hypothesis $p_{ij} = p_{i,} \cdot p_{,j}$, i.e., to the hypothesis of independence of rows and columns, i.e., of no association in the contingency table.

The principle of maximum conditional entropy leads us to the same verbal expression for a null hypothesis, even when the marginal probabilities are not assigned. In this case the null hypothesis is of course composite.

The above null hypothesis is closely connected with the assumption made by Good [13] for the purpose of estimating true population probabilities given a large (sampled) contingency table. The assumption was that the log-association factors $\log(p_{ij}/p_{i,} \cdot p_{,j})$ has initially a normal distribution of mean zero. If this assumption were made when only the $p_{i,}$'s and $p_{,j}$'s were known, the subjective expectations of the first-order interactions (see below) would be zero, and this would again suggest the hypothesis of independence of rows and columns.

Example (v). *m-dimensional population contingency table,* $(p_i) = (p_{i_1,i_2,\cdots i_m})$ *with the totals* $p_{i_1,,\cdots}, p_{,i_2,\cdots}, \cdots, p_{,\cdots,i_m}$ *assigned, i.e., the totals of all complete* $(m - 1)$*-dimensional blocks.*

The restraints are

$$(1) \qquad \sum_{i_2,i_3,\cdots,i_m} p_i = p_{i_1,,\cdots,}, \text{ etc.}$$

By using Lagrange's method of undetermined multipliers, we find equations of the following form for the distribution (p_i) of maximum entropy: $1 + \log p_i = \lambda_{i_1}^{(1)} + \lambda_{i_2}^{(2)} + \cdots + \lambda_{i_m}^{(m)}$. By giving i the two values $(i_1, i_2, \cdots, i_m)$ and $(i_1', i_2, \cdots, i_m)$ in turn, we see that

$$\log (p_{i_1 i_2 \cdots i_m}/p_{i_1' i_2 \cdots i_m}) = \lambda_{i_1}^{(1)} - \lambda_{i_1'}^{(1)}.$$

From this we can deduce that p_i is of the form

$$p_i = g_1(i_1)g_2(i_2) \cdots g_m(i_m),$$

and hence that

$$p_i = p_{i_1,,\cdots,}p_{,i_2,\cdots,} \cdots p_{,,\cdots,i_m}.$$

Thus the principle of maximum entropy generates for consideration the null hypothesis of independence of the coordinates of i. Verbally the same null hypothesis can be generated by the principle of maximum conditional entropy.

In the remainder of this paper we shall be concerned with other natural restraints on multidimensional population contingency tables, and consequently with generalizations of the notion of independence. To say that the rows and columns of a population contingency table are independent is equivalent to saying that the rank of the table, regarded as a matrix, is unity. Some of our work can be interpreted as generalizing the notion of unit rank to more than two dimensions. But see the remark following Theorem 3, part (v).

For an excellent survey of the literature of interaction in multidimensional contingency tables see B. N. Lewis [30]. We cannot agree however with his remark ([30], p. 88) that no new problems arise when one goes beyond three dimensions.

Example (vi). *m-dimensional* $2 \times 2 \times \cdots \times 2$ *population contingency table,* $(p_i) = (p_{i_1 i_2 \cdots i_m})(i_1, i_2, \cdots, i_m = 0, 1)$, *with the subtotals of all rows, columns, shafts, ranks, files, turnpikes, spurs, and corridors assigned,* i.e., *with* $p_{,i_2 \cdots i_m}$, $p_{i_1,i_3 \cdots i_m}$, $\cdots$ $p_{i_1 i_2 \cdots i_{m-1},}$ *assigned.* (All other partial subtotals, namely with more than one acute accent, can be deduced from these.)

In this example there is only one degree of freedom. If (q_i) is a possible set of probabilities with the assigned totals, assuming that there is such a set, then every other set is necessarily of the form $p_i = q_i + (-1)^{|i|}x$, for some real x for which

$$(2) \qquad -\min_{|i| \text{ even}} q_i \leqq x \leqq \min_{|i| \text{ odd}} q_i$$

and where $|i| = i_1 + i_2 + \cdots + i_m$.

We wish to select a null hypothesis by maximising

$$-\sum_{i}^{|i| \text{ even}} (q_i + x) \log(q_i + x) - \sum_{i}^{|i| \text{ odd}} (q_i - x) \log(q_i - x),$$

subject to the above inequality on x. The derivative of this expression with respect to x is

$$-2^{m-1}+2^{m-1}-\sum_{i}^{|i| \text{ even}}\log(q_i+x)+\sum_{i}^{|i| \text{ odd}}\log(q_i-x).$$

By equating this derivative to zero we obtain the condition

$$\prod_{i}^{|i| \text{ even}}(q_i+x)=\prod_{i}^{|i| \text{ odd}}(q_i-x), \tag{3}$$

i.e.,

$$\prod_{i}^{|i| \text{ even}}p_i=\prod_{i}^{|i| \text{ odd}}p_i. \tag{4}$$

The left side of (3) is an increasing function of x, and the right side is a decreasing function of x, so there is at most one root within the permissible range. When $x = -\min q_i(|i| \text{ even})$ the left side is less than the right side, and when $x = \min q_i(|i| \text{ odd})$, the left side exceeds the right side, so there is exactly one root. When $m = 3$, the equation is a cubic in x, and we see that the principle of maximum entropy leads to the same determination of the p_i's as was proposed by Bartlett [1] who described the Condition (4) as that of the vanishing of the second-order interaction for a $2 \times 2 \times 2$ table. The vanishing of second-order interaction can be regarded as an extension of the notion of independence, which, for a three-dimensional table, is the vanishing of both first-order and second-order interactions. (Roy and Kastenbaum [37], describe the vanishing of second-order interaction for a three-dimensional table as "no interaction", but we prefer Bartlett's description.)

Example (vii). *m-dimensional* $2 \times 2 \times \cdots \times 2$ *population contingency table with all "*rth*-order" subtotals assigned.* In order to generalize the above example to some cases where not all the sub-totals are given, it is convenient to introduce some definitions.

If we know the sums of p_i over each subset of $m - r$ coordinates, we say that *we have a complete set of* rth*-order restraints* ($r = 0, 1, \cdots, m-1$). When $r = 0$ we have only the tautological restraint $\sum p_i = 1$. When $r = 1$, we have the case of Example (v), except that we are now thinking mainly in terms of $2 \times 2 \times \cdots \times 2$ tables. When $r = m - 1$, we have the case of Example (vi). If we have a complete set of rth-order restraints then we also have a complete set of all lower-order restraints.

Next we define the m-dimensional mod 2 discrete Fourier transform, a_j^*, of a function a_i of i (see, for example, Good [10]):

$$a_j^*=\sum_{i}(-1)^{ij}a_i \qquad (ij = i_1j_1+\cdots+i_mj_m) \tag{5}$$

for which the inversion formula is

$$a_i=2^{-m}\sum_{j}(-1)^{ij}a_j^*. \tag{6}$$

A complete set of rth-order restraints implies that we have the values of the discrete Fourier transform p_j^* for precisely those j for which $|j| \leqq r$. For moduli greater than 2, this condition would be that the number of non-vanishing components of j must not exceed r. The proof of this assertion is left to the reader.

We define the interactions in the population contingency table by means of the discrete Fourier transform of the logarithms of the probabilities, thus:

$$I_j = \sum_i (-1)^{ij} \log(2^m p_i), \tag{7}$$

and we call I_j an *interaction*, of order $|j| - 1$. The order of an interaction runs from -1 to $m - 1$. In this definition, the factor 2^m automatically drops out if $|j| > 0$.

It should be pointed out that Lancaster [29] and Lewis [30] define three first-order interaction terms for a $2 \times 2 \times 2$ table as the corresponding interactions for the three marginal 2×2 tables. Our meaning for "first-order interaction" is quite different and is more analogous to the definition of interaction in a 2^n-factorial design, which is expressible as an n-dimensional mod 2 discrete Fourier transform [10]. If a new name is required, our interaction could be called a "Fourier interaction" or a "Fourier log-interaction", but, in this paper, we shall call it simply an "interaction" (of appropriate order), and trust that no confusion will arise. For the case $m = 3$, the vanishing of our second-order interaction is equivalent to the vanishing of Bartlett's second-order interaction [1].

We now state our main result for $2 \times 2 \times \cdots \times 2$ tables.

THEOREM 1. *Let (p_i) be an m-way $2 \times 2 \times \cdots \times 2$ population contingency table, with a complete set of* rth-*order restraints. We assume that these restraints are consistent and are not "only just" consistent and so do not force any of the p_i's to vanish. Then the null hypothesis generated by the principle of maximum entropy is that the* rth-*order and all higher-order interactions vanish. If we have no genuine restraints ($r = 0$), then the maximum-entropy null hypothesis is $p_i = 2^{-m}$ (for all i); if $r = 1$, then the maximum-entropy null hypothesis is that of no association, $p_i = p_{i_1,\ldots,}p_{,i_2,\ldots,} \cdots p_{,,\ldots,i_m}$; and if $r = m - 1$, then the maximum-entropy null hypothesis is that the single ($m - 1$)*th-*order interaction vanishes, i.e., that*

$$\prod_i^{|i| \text{ even}} p_i = \prod_i^{|i| \text{ odd}} p_i . \tag{8}$$

PROOF. We wish to maximize $y = -\sum p_i \log p_i$ subject to the restraints and to the inequalities $p_i \geqq 0$ (for all i). These restraints and inequalities define a bounded convex region such that each point on the boundary has at least one of its coordinates, p_i, equal to zero. It can first be proved that the maximum value of y cannot occur on the boundary. Suppose, for example, that there is just one p_i, called p for short, which vanishes. (The following argument generalizes easily to the case of more than one such p.) Let the smallest of the other p_i's be $\delta > 0$. If p is replaced by a small positive number, ϵ, then the restraints can be satisfied by changing some or all of the other p_i's by small quantities

which are proportional to ϵ if ϵ is sufficiently small. (The constant of proportionality will depend on the boundary point.) The total change in the sum of all the terms in y, other than $-p \cdot \log p$, will therefore be less than 2^m times $A\epsilon$, where A does not depend on ϵ. By making ϵ small enough, the change in $p \cdot \log p$ will dominate the situation, since $|\log \epsilon|$ can be made arbitrarily large. Hence the maximum of y cannot occur on the boundary. It must therefore occur at a *stationary* value subject to the restraints. But the Hessian matrix of y is negative definite, so y must have exactly one stationary value, and this must be an interior point of the region of definition.

The restraints can be expressed in the form $\sum_i p_i(-1)^{ij} = p_j^*(|j| \leqq r)$ where the p_j^*'s $(|j| \leqq r)$ are known. We apply Lagrange's method of undetermined multipliers in order to find a stationary point. The Lagrange expression, before differentiation, is $-\sum_i p_i \log p_i + \sum_j \lambda_j p_j^*$, where the λ_j's are the undetermined multipliers, and $\lambda_j = 0$ if $|j| \geqq r$. The partial derivatives with respect to the p_i's are equated to zero, giving the equations $1 + \log p_i = \sum_j \lambda_j(-1)^{ij}$. The right side is seen to be the discrete Fourier transform of (λ_j). If then we take the discrete Fourier transform of both sides, and apply the inversion formula, we see that $I_j = 0$ whenever $|j| > r$.

We can summarize the situation by saying that the maximum is uniquely determined from the equations

$$\sum_i (-1)^{ij} p_i = p_j^*(|j| \leqq r) \tag{9}$$

$$\sum_i (-1)^{ij} \log p_i = 0(|j| > r). \tag{10}$$

It should be mentioned that the problem of numerically maximizing the entropy, subject to an arbitrary set of restraints of the form of knowing sums of p_i's, not necessarily a complete set of rth-order restraints, was treated by David T. Brown [3], following Hartmanis and Lewis [22, 31]. The method is an iterative one in which, at each stage, the "current" values of p_i are scaled so as to satisfy one restraint. The restraints are taken in turn cyclically. We shall refer to the method as the *iterative scaling procedure*. The work of Hartmanis was concerned with the entropy of infinite Markov chains. (Some of its formalism is similar to the likelihood-ratio formalism for a *sample* of a Markov chain in Good [12], which was based on papers by Bartlett and Hoel.)

Example (viii). *m-dimensional* $d_1 \times d_2 \times \cdots \times d_m$ *population contingency table, with all the* r*th-order subtotals assigned.* Let i be a "multipartite residue" mod $(d_1, d_2, \cdots, d_m)$, that is, a "vector" with m components, $i_1, i_2, \cdots, i_m$, where i_1 is a residue mod d_1 and is conveniently represented by one of the integers $0, 1, \cdots, d_1 - 1$, where $i_2, \cdots, i_m$ have similar meanings, and where $d_s \geqq 2\ (s = 1, 2, \cdots, m)$. Let ω_1 be the primitive root of unity, $\omega_1 = \exp[2\pi(-1)^{\frac{1}{2}}/d_1]$, with similar meanings for $\omega_2, \omega_3, \cdots, \omega_m$. *The m-dimensional discrete Fourier* transform mod $(d_1, d_2, \cdots, d_m)$, a_j^*, of a function a_i, of i, is defined by

$$a_j^* = \sum_i \omega_1^{i_1 j_1} \omega_2^{i_2 j_2} \cdots \omega_m^{i_m j_m} a_i = \sum_i \omega^{ij} a_i, \tag{11}$$

where $\omega^{ij} = \omega_1^{i_1 j_1} \cdots \omega_m^{i_m j_m}$, and where j has the same range as i. (See, for example, Good [19].) The Fourier inversion formula is

$$a_i = d_1^{-1} \cdots d_2^{-1} \sum_j \omega^{-ij} a_j^* . \tag{12}$$

Suppose that we have a population contingency table (p_i), and we are given a complete set of rth-order restraints, that is, we know the sums of p_i over each subset of $m - r$ coordinates. It follows that we have the values of the discrete Fourier transform, p_j^*, for precisely those j for which $|j| \leqq r$, where now $|j|$ *is defined as the number of non-vanishing components of* j. We define the (complex) interactions by

$$I_j = \sum_i \omega^{ij} \log (d_1 \cdots d_m p_i), \tag{13}$$

and we call I_j a (complex) interaction of order $|j| - 1$. If $|j| > 0$, the factor $d_1 d_2 \cdots d_m$ is easily seen to be irrelevant.

Although this definition of interaction is a natural extension of (7), it is not of immediate intuitive appeal to a statistician, partly because the interactions are usually complex unless the d_s's are all equal to 2, and partly because the interactions depend on the particular sequence of the rows, and of the columns, and of the corridors, and so on. [The same difficulty can occur when the theory of finite Abelian groups is applied to factorial experiments, as in Fisher [8], and in fact the discrete Fourier transform arises naturally in the representation theory of finite Abelian groups.] The justification of the definition is that it is a useful tool for our purposes, and also that the vanishing of the rth-order and higher-order interactions can be expressed in real forms, as is shown by Theorem 3 below.

We note in passing that if the interaction of order -1 vanishes then all the p_i's must be equal; and also that if all interactions of non-negative order vanish, then again all the p_i's must be equal. The first assertion follows from the fact that $\sum p_i \log p_i$ takes its minimum value when all the p_i's are equal. The second assertion is a simple consequence of the discrete Fourier inversion formula.

THEOREM 2. *Let* (p_i) *be an* m*-dimensional* $d_1 \times d_2 \times \cdots \times d_m$ *population contingency table, with a complete set of* r*th-order restraints. We assume that these restraints are consistent and are not "only just" consistent. Then the null hypothesis generated by the principle of maximum entropy is that the* r*th-order and all higher-order interactions vanish.* (In this paper we shall not be concerned with the possibility that the rth-order interaction vanishes without the vanishing of all higher-order interactions.)

The proof of Theorem 1 can be at once generalized in order to prove Theorem 2, by making use of the more general definition of a discrete multidimensional Fourier transform. It is not necessary to give the details here.

It is interesting to note what would have happened if we had used $\sum p_i^2$ as a measure of "roughness" in place of the negentropy. The application of Lagrange's method of undetermined multipliers would have led to a stationary value of the

roughness at a point where the discrete Fourier transforms of the probabilities themselves, instead of their logarithms, had vanishing components for $|j| > r$. But this stationary point would not necessarily have been possible, since some of its components could have been negative or even complex. The minimum roughness could very well occur on the boundary of the region of possible values of the p_i's, and its determination would be a problem in quadratic planning. Since machine programs exist for quadratic planning, it might be convenient to use the point obtained in this way as a first approximation in an iterative search for the point of maximum entropy, such as the iterative scaling procedure.

The next theorem gives various transformations of the null hypothesis mentioned in Theorem 2. In order to state it we need a definition of a "subtable" of a given table. A subtable is a table obtained by allowing each suffix ("coordinate"), i_s ($s = 1, 2, \cdots, m$), to range only over a subset of its possible values, $0, 1, \cdots, d_s - 1$. A subtable could be "normalized" by dividing all its entries by their sum, but we shall not usually require normalization in this paper. We shall be interested in three kinds of subtables, (i) those in which the dimensionality is reduced below the value m in virtue of one or more of the suffixes being restricted to a single value; (ii) those in which each suffix is restricted to have only two values, so that we get subtables of "size" 2^m; (iii) a mixture of both features, so that we have subtables of size 2^r for $r < m$. It is possible to designate a subtable by stating the subsets to which the indices are restricted; for example, we could talk in a self-explanatory sense of the subtable $(i_1', i_1''; \cdots; i_r', i_r''; i_{r+1}'; i_{r+2}'; \cdots; i_m')$. This could be a subtable of the third of the kinds defined above. If each index is either restricted to a single value or else is not restricted at all, we get a special case of a subtable of the first kind, and call it a *complete subtable* of dimensionality r, or a complete r-way subtable, where r is the number of unrestricted indices.

THEOREM 3. *The following propositions are each equivalent to vanishing of the* rth*-order and all higher-order interactions, i.e., to the null hypothesis of Theorem* 2:

(i) p_i *is a product of* $\binom{m}{r}$ *positive functions, namely a function of* $(i_1, i_2, \cdots, i_r)$ *and a function of* $(i_1, i_2, \cdots, i_{r-1}, i_{r+1})$, *and so on, for all selections of* r *components of* i. (*Compare* (5.1) *of Roy and Kastenbaum* [11].) *For* $r = 0$ *this states that* p_i *is independent of* i.

(ii) *All* rth*-order interactions vanish in all* $(r + 1)$*-dimensional complete subtables.*

(iii) (For the case $r = m - 1$.) *The* $(m - 1)$th*-order interactions vanish in all* $2 \times 2 \times \cdots \times 2$ *subtables of size* 2^m.

(iv) *The* rth*-order and higher-order interactions vanish in all* $2 \times 2 \times \cdots \times 2$ *subtables of size* 2^m. (This last proposition includes (iii) as a special case, but we stated proposition (iii) explicitly for the sake of the intelligibility of the proof below.)

(v) *All* rth*-order interactions vanish in all* $2 \times 2 \times \cdots \times 2$ *subtables of size* 2^{r+1}. *This statement is a natural generalization of the condition for a matrix to be of*

rank 1, *namely that each of its* 2×2 *minors should vanish.* But our definition of "rank 1" is not the same as that in the literature of higher-order determinants (see, for example, L. H. Rice [36]). We are, for example, thinking of a determinant of a $2 \times 2 \times 2$ matrix (a_{ijk}) as defined by the quartic polynomial $a_{000}a_{011}a_{101}a_{110} - a_{100}a_{010}a_{001}a_{111}$, whereas the definitions in the literature are quadratics.

In the following proofs we shall assume $r > 0$, since the theorem is trivial when $r = 0$. We shall refer to the propositions in parts (i) to (v) as "propositions (i) to (v)".

Proof of the necessity and sufficiency of (i). Suppose we are given that the interactions of the rth and higher-orders all vanish, i.e., that $\sum_i \omega^{ij} \log p_i = 0$ if $|j| > r$. Hence, by the inversion formula, $\log p_i = d_1^{-1} \cdots d_m^{-1} \sum_j \omega^{-ij} I_j$ where $I_j = 0$ if $|j| > r$. Hence

$$d_1 d_2 \cdots d_m \log p_i = \sum_{\nu(1),\cdots,\nu(r)}^{1 \leqq \nu(1) < \cdots < \nu(r) \leqq m} \sum_{j_{\nu(1)},\cdots,j_{\nu(r)}}^{j's \neq 0} I_{0\cdots 0 j_{\nu(1)} 0 \cdots 0 j_{\nu(r)} 0 \cdots}$$

$$\omega_{\nu(1)}^{-i_{\nu(1)} j_{\nu(1)}} \cdots \omega_{\nu(r)}^{-i_{\nu(r)} j_{\nu(r)}} \sum_{+\nu(1),\cdots,\nu(r-1)} \sum_{j_{\nu(1)},\cdots,j_{\nu(r-1)}} I_{0\cdots 0 j_{\nu(1)} \cdots 0 j_{\nu(r-1)} 0}$$

$$\ldots 0 \omega_{\nu(1)}^{-i_{\nu(1)} j_{\nu(1)}} \cdots \omega_{\nu(r-1)}^{-i_{\nu(r-1)} j_{\nu(r-1)}} + \cdots = \sum_{\nu(1),\cdots,\nu(r)} F(i_{\nu(1)}, \cdots, i_{\nu(r)})$$

$$+ \sum_{\nu(1),\cdots,\nu(r-1)} F(i_{\nu(1)}, \cdots, i_{\nu(r-1)}) + \cdots$$

where we write F for a function, not usually the same function on each occasion. Therefore

$$\log p_i = \sum_{\nu(1),\cdots,\nu(r)} F(i_{\nu(1)}, \cdots, i_{\nu(r)})$$

where the complex parts of the functions can clearly be removed. Hence

$$p_i = \prod_{\nu(1),\cdots,\nu(r)}^{1 \leqq \nu(1) < \cdots < \nu(r) \leqq m} F(i_{\nu(1)}, \cdots i_{\nu(r)}), \tag{14}$$

where the functions are positive. This proves the necessity part of (i). In order to prove the sufficiency, it is enough, in virtue of the linearity of the interactions as functions of $(\log p_i)$, to suppose that $\log p_i = F(i_1, i_2, \cdots, i_r)$, where we have selected the first r indices merely for the sake of simplicity of notation. When $|j| > r$, one of the components $j_{r+1}, \cdots, j_m$, say j_ν, must be non-zero, so that

$$\sum_{i_\nu} \omega_\nu^{i_\nu j_\nu} \log p_i = 0,$$

and so

$$\sum_i \omega^{ij} \log p_i = 0 \ (|j| > r),$$

as required.

Proof of the necessity and sufficiency of (ii). Write $\log p_i = a_i$. If all the rth-

order and higher-order interactions vanish we have in particular that

$$\sum_{i} a_{i_1 \cdots i_m} \omega_1^{i_1 j_1} \cdots \omega_m^{i_m j_m} = 0, \tag{15}$$

whenever $j_1, \cdots, j_{r+1}$ are all non-zero. Take the $(m - r - 1)$-dimensional inverse discrete Fourier transform with respect to $j_{r+2}, j_{r+3}, \cdots, j_m$, which is possible since these variables are unrestricted, and we get

$$\sum_{i_1, \cdots, i_{r+1}} a_{i_1 \cdots i_m} \omega_1^{i_1 j_1} \cdots \omega_{r+1}^{i_{r+1} j_{r+1}} = 0 \tag{16}$$

$$(j_1 \not\equiv 0, j_2 \not\equiv 0, \cdots, j_{r+1} \not\equiv 0; i_{r+2}, \cdots, i_m \text{ unrestricted}).$$

Conversely Condition (16) implies Condition (15). The same argument could obviously be applied to any subset of $r + 1$ "coordinates", instead of the first $r + 1$, and part (ii) of the theorem follows.

Proof of the sufficiency of (iii) *for the case* $r = m - 1$. Consider a $2 \times 2 \times \cdots \times 2$ subtable of size 2^m, designated by $(i_1, i_1'; \cdots; i_m, i_m')$. For convenience in expressing the vanishing of the $(m - 1)$th-order interaction in this subtable we introduce the notation $i_s(0) = i_s$, $i_s(1) = i_s'$ $(s = 1, 2, \cdots, m)$. Note that we are using the symbol i_s both as a functional symbol and as a particular one of its two values. We have

$$\sum_{\delta_1, \cdots, \delta_m}^{0,1} (-1)^{\delta_1 + \cdots + \delta_m} a_{i_1(\delta_1), \cdots, i_m(\delta_m)} = 0.$$

Consequently

$$\sum_{\delta_1, \cdots, \delta_{m-1}}^{0,1} (-1)^{\delta_1 + \cdots + \delta_{m-1}} a_{i_1(\delta_1), \cdots, i_{m-1}(\delta_{m-1}), i_m} = \sum_{\delta_1, \cdots, \delta_{m-1}}^{0,1} (-1)^{\delta_1 + \cdots + \delta_{m-1}} a_{i_1(\delta_1), \cdots, i_{m-1}(\delta_{m-1}), i_m'}.$$

Since the only distinction between the two sides of this equation is that the suffix i_m on the left is replaced by i_m' on the right, and since i_m' can be any integer from 0 to $d_m - 1$, it follows that the left side is mathematically independent of i_m, and is therefore of the form $F(i_1, i_1', \cdots, i_{m-1}, i_{m-1}')$, where, as before, we use the symbol F to denote a function, not usually the same one on each of its occurrences. Now put $i_1' = \cdots = i_{m-1}' = 0$. We get

$$a_{i_1 \cdots i_m} = F(i_2, \cdots, i_m) + \cdots + F(i_1, \cdots, i_{m-2}, i_m) + F(i_1, \cdots, i_{m-1}). \tag{17}$$

The sufficiency of part (iii) of Theorem 3 now follows from that of part (i).

Proof of the necessity of part (iv) (*and therefore of part* (iii)). We are given that the interactions of the rth and higher orders all vanish. Therefore, proposition (i) is true. *A fortiori* it is true for any assigned $2 \times 2 \times \cdots \times 2$ subtable of size 2^m. Therefore by the sufficiency of (i) *applied to the subtable*, we see that

the rth and higher-order interactions (which are real) in the subtable must vanish. This is true for each subtable.

Another proof of necessity. Let $\epsilon = (\epsilon_1, \cdots, \epsilon_m)$ where

$$\epsilon_\nu = 1 \text{ if } \nu = \nu_1 \text{ or } \nu_2 \text{ or } \cdots \text{ or } \nu_{r+1}$$
$$= 0 \text{ otherwise.}$$

A typical rth-order (real) interaction in the subtable $(i_1, i_1'; \cdots; i_m, i_m')$ is x, defined as

$$x = \sum_{\delta_1,\cdots,\delta_m}^{0,1} (-1)^{\epsilon_1\delta_1+\cdots+\epsilon_m\delta_m} \log p_{i_1(\delta_1),\cdots,i_m(\delta_m)},$$

where, as before, $i_1(0) = i_1, i_1(1) = i_1', \cdots, i_m(0) = i_m, i_m(1) = i_m'$. We wish to prove that $x = 0$. We have

$$x = d_1^{-1} \cdots d_m^{-1} \sum_j I_j(\omega_1^{-i_1 j_1} \pm \omega_1^{-i_1' j_1}) \cdots (\omega_m^{-i_m j_m} \pm \omega_m^{-i_m' j_m})$$

where the factor $\omega_\nu^{-i_\nu j_\nu} \pm \omega_\nu^{-i_\nu' j_\nu}$ has a minus sign if ν is one of $\nu_1, \cdots, \nu_{r+1}$ and has a plus sign otherwise. Hence the coefficient of I_j vanishes if any of $j_{\nu_1}, j_{\nu_2}, \cdots, j_{\nu_{r+1}}$ is 0. But $I_j = 0$ if j has as many as $r + 1$ non-zero components. Therefore $x = 0$, i.e., the rth-order real interactions vanish in any $2 \times 2 \times \cdots \times 2$ subtable of size 2^m.

Proof of the sufficiency of part (iv). For each $2 \times 2 \times \cdots \times 2$ subtable of size 2^m, all interactions of orders r and higher vanish. Therefore, by part (ii) *applied to these subtables,* the rth-order interactions vanish in each $(r + 1)$-dimensional $2 \times 2 \times \cdots \times 2$ subtable of size 2^{r+1}. Therefore, by the sufficiency of part (iii) *with m replaced by $r + 1$,* we see that all rth-order interactions vanish in each *complete* subtable of dimensionality $r + 1$. The sufficiency of part (iv) then follows from that of part (ii).

Part (v) of the theorem is a trivial consequence of parts (ii) and (iv), so that the proof of Theorem 3 is now complete.

We conclude Section 2. with some tests of significance of the null hypothesis.

Tests of significance for no rth-order or higher-order interaction within the wider hypothesis of no sth-order or higher-order interaction. Suppose that we have an ordinary (sampled) contingency table (n_i). The principle of maximal conditional entropy generates the hypotheses of no rth-order or higher-order interaction, for $r = 0, 1, 2, \cdots, m - 1$. Let us call these the hypotheses H_r; for example, H_1 is the hypothesis of independence. For formal convenience we write H_m for the hypothesis that states nothing at all. It is natural now to ask for statistical tests for testing H_r within H_s when $r < s$. The case $r = 1, s = m$ is classical, and the case $r = m - 1, s = m$, $m = 3$, is the one treated by Bartlett [1] and Roy and Kastenbaum [37].

We shall consider the chi-squared test and the likelihood-ratio test. For the application of either of these tests we must first maximize the likelihood, or equivalently maximize $z = \sum_i n_i \log p_i$, subject to the restraints $I_j = 0, |j| > r$, and $\sum_i p_i = 1$. We now have a Lagrange expression of the form

$$\sum_i n_i \log p_i - \sum_j \lambda_j \sum_i \omega^{ij} \log (d_1 \cdots d_m p_i) - \lambda \sum p_i$$

where this time λ_j is defined as zero if $|j| \leqq r$, the opposite of what was done in Example (vii). The conditions for stationarity are

$$(n_i/p_i) = \lambda + \sum_j \lambda_j \omega^{ij}/p_i = \lambda + \lambda_i^*/p_i .$$

This gives us $\lambda p_i = n_i - \lambda_i^*$, $\lambda = n_{,,\ldots,} - d_1 \cdots d_m \lambda_0 = n_{,,\ldots,} = n$ say, $\lambda p_j^* = n_j^* - d_1 \cdots d_m \lambda_j$, and hence $p_j^* = n_j^*/n$ when $|j| \leqq r$. These equations then must be satisfied by the maximum-likelihood value of (p_i), subject to no rth-order or higher-order interaction, provided that the likelihood is maximised at a stationary value.

If we write more explicitly, $\hat{p}_i^{(r)}$, we see that the rth-order sums of the $n\hat{p}_i^{(r)}$'s are equal to the rth-order sums of the n_i's. (See the remark following Equation (6).) This condition is so simple that it might well be capable of a simpler proof. At any rate, it is necessary to adjoin to it the vanishing of the rth and higher-order interactions, in order that the $\hat{p}_i^{(r)}$'s should be determined.

By combining this result with Theorem 2, we have a curious duality connecting maximum likelihood with maximum entropy:

THEOREM 4. *The maximum-likelihood values of p_i, namely $\hat{p}_i^{(r)}$, subject to the vanishing of the* rth *and all higher-order interactions, $I_j = 0$ $(|j| > r)$, are equal to the maximum-entropy values of p_i, subject to the* rth*-order sums of the np_i's being equal to those of the n_i's, provided that the maximum likelihood is reached at a stationary value of the likelihood.*

The numerical problem of solving a set of equations, some linear in the independent variables and some linear in their logarithms, is one whose optimal solution is by no means obvious. The work on the iterative numerical solution of general sets of equations (see, for example, [21], [38], [43]) could be used; but, in view of Theorem 4, our maximum-likelihood equations could be solved by means of the iterative scaling procedure for maximizing entropy. (See remarks following Equation (10).)

For the case $r = 2$, $m = s = 3$, i.e., for the testing of no second-order interaction in a three-dimensional table, our $\hat{p}_i^{(2)}$ agrees, for example, with those in [37]. In order to find $\hat{p}_i^{(2)}$, Roy and Kastenbaum [37] needed to solve $(d_1 - 1)$ $(d_2 - 1)$ $(d_3 - 1)$ simultaneous non-linear equations in as many unknowns, Darroch [5] had $d_2d_3 + d_3d_1 + d_1d_2$ equations, and Kastenbaum and Lamphiear [25] produced an iterative procedure for solving the $(d_1 - 1)$ $(d_2 - 1)$ $(d_3 - 1)$ equations, by extending a method used by Norton [33]. Any one of these methods could be used for solving our $d_1d_2d_3$ equations. For other values of r, s, and m, iterative methods could presumably be devised for solving our equations, or else Theorem 4 could be invoked and then an obvious generalization of the iterative scaling procedure could be applied, along the same lines as Brown [3], who however discusses only the case $d_1 = d_2 = \cdots = d_m = 2$. The matter requires further thought and numerical calculation, possibly rather extensive.

Let

$$\mu_r = -2 \sum n_i \log \hat{p}_i^{(r)}$$

and

$$\chi_r^2 = \sum_i (n_i - n\hat{p}_i^{(r)})^2 / n\hat{p}_i^{(r)}.$$

For testing H_r within H_s $(r < s)$, the Neyman-Pearson likelihood-ratio statistic (see, for example, Wilks [41], p. 419) is $\mu_r - \mu_s$, and asymptotically has a tabular chi-squared distribution having as its number of degrees of freedom the sum of the products of the numbers $d_1 - 1, d_2 - 1, \cdots, d_m - 1$, taken $(r + 1)$ at a time, plus the sum of the products taken $(r + 2)$ at a time, $\cdots$, plus the sum of the products taken s at a time. In order to see that this is the number of degrees of freedom we note that, for example, the value of $p_{0i_2i_3\cdots i_r,,\cdots,,}$ can be deduced from the values of $p_{i_1i_2i_3\cdots i_r,,\cdots,,}$, $(1 \leqq i_1 \leqq d_1 - 1)$ and of $p_{,i_2i_3\cdots i_r,,\cdots,,}$; and that the value of $p_{00i_3\cdots i_r,,\cdots,,}$ can be deduced from those of

$$p_{0i_2i_3\cdots i_r,\cdots,}, p_{i_10i_3\cdots i_r,\cdots,}, p_{,,i_3\cdots i_r,\cdots,}, (1 \leqq i_1 \leqq d_1 - 1, 1 \leqq i_2 \leqq d_2 - 1),$$

and so on. If then we are given all restraints of order $r - 1$, the restraints of order r are determined by

$$p_{,\cdots i_{\nu(1)},\cdots,i_{\nu(2)},\cdots,\cdots,\cdots,i_{\nu(r)},\cdots,\cdots,}$$

where none of $i_{\nu(1)}, i_{\nu(2)}, \cdots, i_{\nu(r)}$ is zero, and where $1 \leqq \nu(1) < \nu(2) < \cdots < \nu(r) \leqq m$. If $d_1 = d_2 = \cdots = d_m = 2$, the number of degrees of freedom is

$$\binom{m}{r+1} + \binom{m}{r+2} + \cdots + \binom{m}{s}.$$

The chi-squared statistic is $\chi_r^2 - \chi_s^2$ and has asymptotically the same distribution as the likelihood-ratio statistic. Note that, for the application of the likelihood-ratio test, tables of $2n \log n$ are useful. A table which I had computed for this purpose, for $n = 1(1)10{,}000$, is included in [28]. The other Ku^3 paper [27] is also somewhat complementary to the present paper.

3. Further Discussions.

New relevant papers. After this paper was submitted for publication, two relevant papers appeared (Plackett [35], Darroch [5]) dealing mainly with three-dimensional contingency tables. Plackett, following Woolf [42], gives a chi-squared test for zero second-order interaction in a $2 \times 2 \times t$ table. It is easier to compute than our chi-squared statistic, but it is not known which tends faster to its asymptotic distribution. Darroch gives the likelihood ratio test for no second-order interaction in a three-dimensional table, and is consistent with the test given above. Note also that Section 3 of [5] contains the conjecture that if, in a three-dimensional table, we have a complete and consistent set of second-order restraints (i.e., all marginal totals), then the hypothesis of no second-order interaction leads to a unique set of cell probabilities (p_i). Provided that the restraints are not "only just consistent", the truth of this conjecture follows from our Theorem 2, combined with the proof of Theorem 1, and with part (iii) of Theorem 3. In fact, in an m-dimensional table with a complete set of rth-order restraints, the hypothesis of the vanishing of the rth and all higher-

order interactions leads to a unique set of cell probabilities, when the restraints are consistent but not only just consistent.

Darroch states that there is little practical interest in interactions of higher than the second order. This may be less true now that there is a definition available. For any given (sampled) contingency table in four or more dimensions, we are now in a position to test the hypothesis of the vanishing of higher-order interactions, and it would be dangerous to assume that they vanish without applying a test. If the tests do not reach significance, owing to the smallness of the sample, or otherwise, then one might be prepared to accept the hypothesis of the vanishing of the rth and higher-order interactions. Then one could apply the inverse discrete Fourier transform (see Equations (13) and (12)) in order to make estimates of the p_i's, which would probably be better than the maximum-likelihood estimates. The corresponding smoothing idea was proposed for factorial experiments by Good [10].

Relationship between analysis of interactions and latent class analysis. L. J. Savage suggests,[1] in a private communication, that the analysis of interactions in contingency tables, as described in the present paper, might be regarded as an alternative to latent class analysis, in that both forms of analysis attempt to approximate expression of multidimensional contingency tables by means of relatively few parameters.

Relationships between maximum entropy, interactions, and Markov chains. Let p_i be the population frequency (probability) of the m-plet i in a stationary and ergodic Markov process of order $m - 1$, with discrete time and a finite alphabet of d letters. The numbers p_i can be regarded as entered into an m-dimensional population contingency table having marginal totals of various dimensionalities less than m. Some of these marginal totals will be probabilities of μ-plets with $\mu < m$; others will be probabilities of what we shall call *split* μ-plets, such as $p_{i_1,i_3,,i_6 i_7 \cdots i_{\mu+3}}$, where there are precisely μ suffixes which have not been summed out, i.e., which are not denoted by acute accents. The marginal totals are distinguished from those of a general m-dimensional population contingency table in that there are additional conditions of consistency, namely that $p_{,i_2 \cdots i_m} = p_{i_2 \cdots i_{m'}}$, for all $i_2, \cdots, i_{m'}$. If the process were of order only $m - 2$, the marginal totals would satisfy the further consistency conditions, relating to the split $(m - 1)$-plet population frequencies:

$$p_{i_1,i_3 \cdots i_m} = \sum_{i_2} (p_{i_1 \cdots i_{m-1}} p_{i_2 \cdots i_m} / p_{i_2 \cdots i_{m-1}})$$

etc. We can then derive the following result:

THEOREM 5. *Given an ergodic process with discrete time and a finite alphabet, and given that its $(m - 1)$-plet and split $(m - 1)$-plet population frequencies are consistent with the hypothesis of Markovity of order $m - 2$, then the hypothesis generated by the principle of maximum entropy is indeed that the process is of order $m - 2$.*

---

[1] Both he and L. A. Goodman made a number of other useful comments concerning the manuscript, which I have taken into account.

PROOF. Consider the hypothesis that

$$p_{i_1\cdots i_m} = p_{i_1\cdots i_{m-1}}\, p_{i_2\cdots i_m}/p_{i_2\cdots i_{m-1}},$$

for all $i_1, \cdots, i_m$. These m-plet probabilities form a set which is consistent with the given $(m-1)$-plet probabilities, if these *are* given. Moreover, by part (i) of Theorem 3 this hypothesis is the one that is generated by the principle of maximum entropy applied to the m-dimensional table. We can now apply the same method to $(m+1)$-plets, $(m+2)$-plets, $\cdots$ in turn, and we obtain the null hypothesis that *all* m'-plets, for $m' \geqq m$, have the probabilities required in the statement of the theorem.

Strictly speaking, Theorem 5 should have been stated separately for the cases where the $(m-1)$-plets frequencies are given or are not given. When they are not given it is the principle of maximum *conditional* entropy which is strictly relevant.

Presumably Theorem 5 would remain true for an enumerable, or even for a non-enumerable alphabet.

The naturalness of our definition of interaction is supported by Theorem 6 below.

Let the *span* of a vector be defined as the largest number of consecutive components such that all the components to the left and right of these vanish. For example, a zero vector has span zero whatever its dimensionality, and the span of (0, 0, 5, 0, 1, 0, 0) is 3. Then we have:

THEOREM 6. *If a process is Markovian of order* $m-2$, *then* $I_j = 0$, *whenever the span of* j *is* m *or more.*

PROOF. We define the transition probability $q_{i_1\cdots i_{m-1}}$ as the probability that a letter is i_{m-1} given that the previous $(m-2)$-plet was $(i_1, \cdots, i_{m-2})$. Let

$$a_{i_1\cdots i_{m-2}} = \log p_{i_1\cdots i_{m-2}},\ b_{i_1\cdots i_{m-1}} = \log q_{i_1\cdots i_{m-1}}.$$

We have

$$p_{i_1 i_2\cdots i_{m+l}} = p_{i_1\cdots i_{m-2}}\, q_{i_1\cdots i_{m-1}}\, q_{i_2\cdots i_m} \cdots q_{i_{l+2}\cdots i_{m+l}},$$

whenever $l \geqq 0$. It follows that, if $\omega = \exp(2\pi i/d)$,

$$\begin{aligned} I_{j_1\cdots j_{m+l}} &= \sum_i \{(a_{i_1\cdots i_{m-2}} + b_{i_1\cdots i_{m-1}} + b_{i_2\cdots i_m} + \cdots + b_{i_{l-2}\cdots i_{m+l}}) \\ &\qquad \omega^{i_1 j_1 + \cdots + i_{m+l} j_{m+l}}\} \\ &= \sum_{i_{m-1}\cdots i_{m+l}} \omega^{i_{m-1} j_{m-1} + \ldots + i_{m+l} j_{m+l}} \sum_{i_1\cdots i_{m-2}} a_{i_1\cdots i_{m-2}}\, \omega^{i_1 j_1 + \cdots + i_{m-2} j_{m-2}} \\ &\quad + \sum_{i_m\cdots i_{m+l}} \omega^{i_m j_m + \cdots + i_{m+l} j_{m+l}} \sum_{i_1\cdots i_{m-1}} b_{i_1\cdots i_{m-1}}\, \omega^{i_1 j_1 + \cdots + i_{m-1} j_{m-1}} \\ &\quad + \sum_{i_1, i_{m+1}, \cdots, i_{m+l}} \omega^{i_1 j_1 + i_{m+1} j_{m+1} + i_{m+2} j_{m+2} + \cdots + i_{m+l} j_{m+l}} \\ &\qquad \cdot \sum_{i_2, \cdots, i_m} b_{i_2\cdots i_m}\, \omega^{i_2 j_2 + \ldots + i_m j_m} + \cdots + \sum_{i_1\cdots i_{l+1}} \omega^{i_1 j_1 + \cdots + i_{l+1} j_{l+1}} \\ &\qquad\qquad \cdot \sum_{i_{l+2}\cdots i_{m+l}} b_{i_{l+2}\cdots i_{m+l}}\, \omega^{i_{l+2} j_{l+2} + \cdots + i_{m+l} j_{m+l}}, \end{aligned}$$

and this vanishes if neither j_1 nor j_{m+l} is zero.

Theorem 6 is true even if $m = 1$ or 2, provided that "Markovity of order 0" is interpreted to mean 'random' (independence), and "Markovity of order -1" to mean "perfectly random" or "flat-random" (independence and equiprobability).

I have not yet proved the converse of Theorem 6.

Contingency table with a small sample. Deming *et al* ([40], [6]) considered the problem of estimating population frequencies in a contingency table when the marginal population frequencies, $(p_{i,})$ and $(p_{,j})$ are known, and at the same time there is a small sample (n_{ij}) of the interior, where $\sum_{ij} n_{ij} = N$. In the present paper we have assumed the sample size N to be zero when considering contingency tables. They proposed a least squares method, in which

$$\sum_{i,j} (p_{ij} - n_{ij}/N)^2/n_{ij}$$

was minimised subject to the restraints

$$\sum_j p_{ij} = p_{i,} , \sum_i p_{ij} = p_{,j} .$$

They also considered extensions to multidimensional contingency tables. Although their method has been useful for census work, it is clear that some modification is required when some of the n_{ij}'s vanish. Their method would lead to $p_{ij} = 0$ for such n_{ij}'s, whereas ours, with $N = 0$, leads to $p_{ij} = p_{i,}p_{,j}$. The method of Good [13], for the case where $N \neq 0$, would lead to a compromise between these two estimates.

Another possible method of effecting this compromise, perhaps applicable to multidimensional tables, could be obtained by combining maximum likelihood with maximum entropy. The latter refers to the situation where we have an infinite sample for some marginal distributions and zero sample for the complete distribution. When there is a non-zero sample for the complete distribution, it is reasonable first to test the null hypotheses that are thrown up by the principle of maximum entropy. If these are rejected, then we are faced with an estimation problem.

A possible approach, which however is difficult to justify, would be to maximize some linear combination of the entropy and of the log-likelihood, such as the sum. For a two-dimensional contingency table, with assigned marginal population frequencies, this leads to the rule: *maximize*

$$\sum (n_{ij} - p_{ij}) \log p_{ij} ,$$

subject to the restraints. The result would presumably be to estimate p_{ij} somewhere between $p_{i,}p_{,j}$ and n_{ij}/n, and in this respect would resemble the methods of Good [13].

The above method is equivalent to the selection of the distribution of maximum final credibility density, assuming the initial density proportional to the exponential of the entropy.

This rule for an initial credibility density is not consistent, for a multinomial distribution, with rules that have been previously advocated, since it gives a

density proportional to $\prod p_i^{-p_i}$; whereas previous rules have been of the form $\prod p_i^{\alpha}$ ([34], p. 308, and references in [15], p. 862). Previous suggestions have not been mutually consistent. One of the most interesting is Jeffreys' invariance rule [24], for which $\alpha = -\frac{1}{2}$. We conclude this paper with a further comment concerning this rule.

Another possible method of generating null hypotheses. Jeffreys' invariance rule assigns a possible credibility distribution to a parameter space. It would be reasonable to think of this credibility distribution as appropriate for the non-null hypothesis and to select null hypotheses at stationary values of the invariant density. For the simplest example, a multinomial distribution, the null hypothesis generated by this principle of stationary invariant density is the "equiprobable" or "flat" multinomial distribution; but I suspect that this principle seldom throws up the same null hypotheses as the principle of maximum entropy; for example, it does not do so for a 2×2 contingency table.

Orthogonal and unitary interactions. It is possible to use matrices more general than (ω^{ij}) in order to define interactions with many of the properties of I_j. This is exemplified by the following duality theorem which generalizes Theorem 4. Its proof is omitted in order to save space.

THEOREM 7. *Let (ω_{ij}) be proportional to a symmetric unitary matrix (or to a real symmetric orthogonal matrix) for which $\omega_{0j} = 1$ for all j, and let*

$$\mathcal{J}_j = \sum_i \omega_{ji} \log p_i .$$

Let R be a class of values of j containing $j = 0$. Then the maximum-likelihood values of p_i, subject to $\mathcal{J}_j = 0$ when $j \notin R$, satisfy the equations $np_j^{\ddagger} = n_j^{\ddagger}$ when $j \in R$, where $p_j^{\ddagger} = \sum_i \omega_{ji} p_i$, $n_j^{\ddagger} = \sum_i \omega_{ji} n_i$. Dually, the maximum entropy values of p_i, subject to $np_j^{\ddagger} = n_j^{\ddagger}$ for $j \in R$ satisfy the equations $\mathcal{J}j = 0$ when $j \notin R$.

Further; the equations $np_j^{\ddagger} = n_j^{\ddagger}$ when $|j| \leqq r$ are equivalent to the proposition that the rth order sums of the np_i's are equal to those of the n_i's, provided that (i) *ω_{ij} is a function of the m products $i_1 j_1, i_2 j_2, \cdots, i_m j_m$; and* (ii) *$(\omega_{ij})$ is proportional to a unitary matrix.*

The possibility of generalizing our original definition of Fourier interactions was suggested by L. A. Goodman (private communication) although not in connection with the duality theorem.

REFERENCES

[1] BARTLETT, M. S. (1935). Contingency table interactions. *J. Roy. Statist. Soc. Suppl.* **2** 248–252.

[2] BOLTZMANN, L. (1877). *Wiener Sitzungsberichte* **76** 373 (cited in [7], p. 30).

[3] BROWN, DAVID T. (1959). A note on approximations to discrete probability distributions. *Information and Control* **2** 386–392.

[4] COX, D. R. (1962). Further results on tests of separate families of hypotheses. *J. Roy. Statist. Soc. Ser. B* **24** 406–424.

[5] DARROCH, J. N. (1962). Interactions in multi-factor contingency tables. *J. Roy. Statist. Soc. Ser. B* **24** 251–263.

[6] DEMING, W. EDWARDS and STEPHAN, FREDERICK W. (1940). On a least squares adjustment of a sampled frequency table. *Ann. Math. Statist.* **11** 427–444.

[7] EHRENFEST, P. and EHRENFEST, T. (1959). *The Conceptual Foundations of the Statistical Approach in Mechanics* (English translation by M. J. Maravcsik from a German article of 1912). Cornell Univ. Press.

[8] FISHER, R. A. (1942). The theory of confounding in factorial experiments in relation to the theory of groups. *Ann. Eugenics*, **11** 341–353.

[9] GOOD, I. J. (1953). Contribution to the discussion of a paper by M. S. Bartlett, in *Symposium on Information Theory*, Ministry of Supply, London (1950), reprinted in *Trans. I.R.E.*, IT 180–181.

[10] GOOD, I. J. (1958). The interaction algorithm and practical Fourier analysis. *J. Roy. Statist. Soc. Ser. B* **20** 361–372 and addendum **22** (1960) 372–375.

[11] GOOD, I. J. (1950) *Probability and the Weighing of Evidence*. Griffin, London.

[12] GOOD, I. J. (1955). The likelihood ratio test for Markoff chains. *Biometrika* **42** 531–533 and correction **44** 301.

[13] GOOD, I. J. (1955). On the estimation of small frequencies in contingency tables. *J. Roy. Statist. Soc. Ser. B* **18** 113–124.

[14] GOOD, I. J. (1956). Some notation and terminology in information theory. *Proc. Inst. Elec. Engrs. C* **103** 200–204.

[15] GOOD, I. J. (1957). Saddle-point methods for the multinomial distribution. *Ann. Math. Statist.* **28** 861–881.

[16] GOOD, I. J. (1959). Kinds of probability. *Science* **129** 443–447.

[17] GOOD, I. J. (1960). Contribution to the discussion of a paper by E. M. L. Beale. *J. Roy. Statist. Soc. Ser. B* **22** 79–82.

[17a] GOOD, I. J. (1961). The multivariate saddle-point method and chi-squared for the multinomial distribution. *Ann. Math. Statist.* **32** 535–548.

[18] GOOD, I. J. (1962). Contribution to the discussion of a paper by C. Stein. *J. Roy. Statist. Soc. Ser. B* **24** 289–291.

[19] GOOD, I. J. (1962). Analogues of Poisson's summation formula. *Amer. Math. Monthly* **69** 259–266.

[20] GOOD, I. J. and DOOG, K. CAJ (1958 and 1959). A paradox concerning rate of information. *Information and Control* **1** 113–126 and **2** 195–197. (Continued in **3** (1960) 116–140 by the senior author.)

[21] HART, WILLIAM L. and MOTZKIN, THEODORE S. (1956). A composite Newton-Raphson gradient method for the solution of systems of equations. *Pacific J. Math.* **6** 691–707.

[22] HARTMANIS, J. (1959). The application of some basic inequalities for entropy. *Information and Control* **2** 199–213.

[23] JAYNES, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.* **106** 620–630.

[24] JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser. A* **186** 453–461.

[25] KASTENBAUM, M. A. and LAMPHIER, D. E. (1959). Calculation of chi-square to test the no three-factor interaction hypothesis. *Biometrics* **15** 107–115.

[26] KULLBACK, SOLOMON (1959). *Information Theory in Statistics*. Wiley, New York.

[27] KULLBACK, S., KUPPERMAN, M. and KU, H. H. (1962). An application of information theory to the analysis of contingency tables, with a table of $2n \ln n$, $n = 1(1)$, 10,000. *J. Res. Nat. Bur. Standards* **66 B** 217–243.

[28] KULLBACK, S., KUPPERMAN, M. and KU, H. H. (1962). Tests for contingency tables and Markov chains. *Technometrics* **4** 573–608.

[29] LANCASTER, H. O. (1951). Complex contingency tables treated by the partition of χ^2. *J. Roy. Statist. Soc. Ser. B* **13** 242–249.

[30] LEWIS, B. N. (1962). On the analysis of interaction in multidimensional contingency tables. *J. Roy. Statist. Soc. Ser. A* **125** 88–117.

[31] LEWIS, P. M., II (1959). Approximating probability distributions to reduce storage requirements. *Information and Control* **2** 214–225.

[32] LINDLEY, D. V. (1961). The use of prior probability distributions in statistical inference and decision. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 453–468. Univ. of California Press.

[33] NORTON, H. W. (1945). Calculation of chi-square for complex contingency tables. *J. Amer. Statist. Assoc.* **40** 251–258.

[34] PERKS, WILFRED (1947). Some observations on inverse probability including a new indifference rule. *J. Inst. Actuar.* **73** 285–312 and discussion 313–334.

[35] PLACKETT, R. L. (1962). A note on interactions in contingency tables. *J. Roy. Statist. Soc. Ser. B* **24** 162–166.

[36] RICE, L. H. (1930). Introduction to higher determinants. *J. Math. Physics* **9** 47–71.

[37] ROY, S. N. and KASTENBAUM, MARVIN A. (1956). On the hypothesis of no 'interaction' in a multi-way contingency table. *Ann. Math. Statist.* **27** 749–757.

[38] SCARBOROUGH, JAMES B. (1958). *Numerical Mathematical Analysis.* Johns Hopkins Press, Baltimore.

[39] SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423 and 623–656.

[40] STEPHAN, FREDERICK F., DEMING W. EDWARDS and HANSEN, MORRIS H. (1940). The sampling procedure of the 1940 population census. *J. Amer. Statist. Assoc.* **35** 615–630.

[41] WILKS, S. S. (1962). *Mathematical Statistics.* Wiley, New York. (See p. 419).

[42] WOOLF, B. (1955). On estimating the relation between blood group and disease. *Ann. Human Genetics* **19** 251–253.

[43] ZAGUSKIN, V. L. (1961). *Handbook of Numerical Methods for the Solution of Algebraic and Transcendental Equations* (trans. from the Russian by G. O. Harding). Oxford Univ. Press.