*The Annals of Mathematical Statistics*
1970, Vol. 41, No. 1, 214–226

# THE WEIGHTED LIKELIHOOD RATIO, SHARP HYPOTHESES ABOUT CHANCES, THE ORDER OF A MARKOV CHAIN

BY JAMES M. DICKEY[1] AND B. P. LIENTZ

*State University of New York at Buffalo and System Development Corporation*

**1. Summary.** The Bayesian theory for testing a sharp hypothesis, defined by fixed values of parameters, is here presented in general terms. Arbitrary positive prior probability is attached to the hypothesis. The ratio of posterior to prior odds for the hypothesis is given by the weighted likelihood ratio, shown here to equal Leonard J. Savage's (1963) ratio of a posterior to a prior density (2.21). This Bayesian approach to hypothesis testing was suggested by Jeffreys (1948), Savage (1959), (1961), Lindley (1961), and Good (1950), (1965), but obscured some what by approximations and unique choices of prior distributions. This Bayesian theory is distinct from that of Lindley (1965) and that of Dickey (1967a).

Applications are given to hypotheses about multinomial means, for example, equality of two binomial probabilities. A new test is presented for the order of a finite-state Markov chain.

**2. Introduction.** Assume a statistical model in which the observed data vector $\mathbf{D} \in E^n$ occurs according to the probability mass or density function (elementary derivative) $\varphi(\mathbf{D} \mid \theta)$, depending continuously on an unknown parameter vector $\boldsymbol{\theta} \in E^r$. Assume an individual's opinion about $\boldsymbol{\theta}$ before and after his observation of $\mathbf{D}$ is described by his prior and posterior probability distributions $P(S)$ and $P(S \mid \mathbf{D})$. We thus take $\mathbf{D}$ and $\boldsymbol{\theta}$ to have a well-behaved joint distribution.

Suppose one suspects the unknown parameter $\boldsymbol{\theta}$ of belonging to a given Borel set $H \subset E^r$. Then $0 < P(H) < 1$, and hence, except for mathematical pathologies, $0 < P(H \mid \mathbf{D}) < 1$.

Let $\bar{H}$ denote a Borel measurable "alternative": $H \cap \bar{H} = \varnothing$ with $P(H) + P(\bar{H}) = 1$. Let $O$ denote odds corresponding to probabilities $P$,

$$O(H) = P(H)/P(\bar{H}), \tag{2.1}$$

having the immediate properties, $O(\bar{H}) = 1/O(H)$ and $P(H) = O(H)/[1 + O(H)]$.

Denote the posterior odds,

$$O(H \mid \mathbf{D}) = P(H \mid \mathbf{D})/P(\bar{H} \mid \mathbf{D}); \tag{2.2}$$

and the ratio of posterior to prior odds for $H$,

$$L = O(H \mid \mathbf{D})/O(H). \tag{2.3}$$

Received November 12, 1968; revised June 3, 1969.

[1] Research partially supported by Public Health Service Research Grants No. CA-10801 from the National Cancer Institute and No. GM-16557 from the Institute of General Medical Sciences.

COROLLARY (Well Known).

$$L = \Phi(\mathbf{D} \mid H)/\Phi(\mathbf{D} \mid \bar{H}), \quad \text{where} \tag{2.4}$$

$$\Phi(\mathbf{D} \mid H) = \int \varphi(\mathbf{D} \mid \boldsymbol{\theta})\, dP(\boldsymbol{\theta} \mid H), \tag{2.5}$$

$$\Phi(\mathbf{D} \mid \bar{H}) = \int \varphi(\mathbf{D} \mid \boldsymbol{\theta})\, dP(\boldsymbol{\theta} \mid \bar{H}). \tag{2.6}$$

PROOF. Since, if $\varphi(\mathbf{D} \mid \boldsymbol{\theta})$ is a density, it is assumed to be a uniquely defined elementary derivative (and hence so is $\Phi(\mathbf{D} \mid H)$), then Bayes' theorem applies.

$$P(H \mid \mathbf{D}) = P(H)\Phi(\mathbf{D} \mid H)/\Phi(\mathbf{D}), \quad \text{where} \tag{2.7}$$

$$\Phi(\mathbf{D}) = P(H)\Phi(\mathbf{D} \mid H) + P(\bar{H})\Phi(\mathbf{D} \mid \bar{H}); \tag{2.8}$$

and analogously for $P(\bar{H} \mid \mathbf{D})$. Substitution of (2.7) and its analog into (2.3) yields (2.4).

In case $H$ and $\bar{H}$ are simple point hypotheses, $L$ is the usual likelihood-ratio-test statistic. More generally, Wald (1947), Jeffreys (1948), Good (1950), (1965), Lindley (1961), Savage (1959), (1961), in effect Raiffa and Schlaifer (1961), and Barnard (1964) have proposed the use, in inference about $H$ from $\mathbf{D}$, of the ratio $L$ of weighted averages $\Phi$ of $\varphi$ over $H$ and $\bar{H}$. To use a weighted likelihood ratio, to summarize the evidence in $\mathbf{D}$ for $H$, is certainly more sound in principle than the usual likelihood ratio, involving a constrained and an unconstrained maximum of $\varphi$.

Following Wald, Raiffa and Schlaifer and others have taken a general decision-theoretic view of testing. Certain such concepts of tests for linear hypotheses within a context of prediction or estimation were proposed by Dickey (1967a) and by Lindley (1968). Non-decision-theoretic, whether or not called Bayesian (as in Lindley (1965)), usually involve tail areas and should be avoided when genuine Bayesian tests are available. Good (1965) writes of "Bayes/non-Bayes compromises," in which tail probabilities are developed for the statistic $L$. A Bayesian test does not, of course, depend on the (noninformative) stopping rule.

We restrict attention here to a "sharp hypothesis," defined as follows. Given an invertible transformation $\boldsymbol{\xi}$ having non-zero Jacobian,

$$\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\xi}), \tag{2.9}$$

partition the vector $\boldsymbol{\xi} \in E^r$,

$$\boldsymbol{\xi} = (\boldsymbol{\eta}', \boldsymbol{\zeta}')', \qquad \boldsymbol{\eta} \in E^s, \qquad s \leqq r. \tag{2.10}$$

We seek to test the possibly composite hypothesis

$$H: \boldsymbol{\eta} = \boldsymbol{\eta}_0, \tag{2.11}$$

where $\boldsymbol{\eta}_0$ is a fixed constant, against the alternative

$$\bar{H}: \boldsymbol{\eta} \neq \boldsymbol{\eta}_0. \tag{2.12}$$

We take special interest in linear $H$,

$$\boldsymbol{\xi} = \boldsymbol{\Lambda}\boldsymbol{\theta}.$$

By an abuse of notation, denote again by $P$ the induced prior measure for $\xi$. Assume as the support of $P$ an analytic surface segment $\Xi \subset E^r$ of dimension $\dim(\Xi)$, where $\dim(H \cap \Xi) < \dim(\Xi) \leqq r$. We shall use integration with respect to Lebesgue measure $\mu$ on the surface $\Xi$ with differential element denoted $d\xi = d\zeta\, d\eta$; and then on the measure-zero set $H \cap \Xi$, integration with respect to the factor Lebesgue measure $\mu_1$ with element $d\zeta$.

The prior probability measure is a mixture over $H$ and $\bar{H}$ of the following assumed form. For a set $S \subset \Xi^r$ with Borel intersection $S \cap \Xi$ (and hence $S \cap \Xi \cap H$),

$$P(S) = P(\bar{H}) \int\int_{S \cap \Xi} f(\eta, \zeta)\, d\zeta\, d\eta + P(H) \int_{S \cap \Xi \cap H} g(\zeta)\, d\zeta. \tag{2.13}$$

The density function $f$ is assumed uniquely defined throughout $\Xi$ as the elementary derivative,

$$f(\xi) = \lim_{\rho \to 0} P(S_\rho(\xi) \mid \bar{H}) / \mu(S_\rho(\xi) \cap \Xi), \tag{2.14}$$

where $S_\rho(\xi)$ denotes the ball of radius $\rho$ centered at $\xi$. Hence, even though for $\xi \in \bar{H}$ $f(\xi)$ has an interpretation as a conditional density of $\xi$ given $\bar{H}$, if $\xi \in H$ $f(\xi)$ is well defined and not necessarily zero.

Similarly, $g$ is assumed given by

$$g(\zeta) = \lim_{\rho \to 0} P(S_\rho(\eta_0, \zeta) \mid H) / \mu_1(S_\rho(\eta_0, \zeta) \cap \Xi \cap H); \tag{2.15}$$

and for $\xi \in H$ $g(\zeta)$ has an interpretation as a conditional density of $\zeta$ given $H$.

For a sharp hypothesis, (2.5) and (2.6) take the forms

$$\Phi(\mathbf{D} \mid H) = \int \varphi(\mathbf{D} \mid \eta_0, \zeta) g(\zeta)\, d\zeta, \tag{2.16}$$

$$\Phi(\mathbf{D} \mid \bar{H}) = \int\int \varphi(\mathbf{D} \mid \eta, \zeta) f(\eta, \zeta)\, d\zeta\, d\eta, \tag{2.17}$$

where $\varphi(\mathbf{D} \mid \eta, \zeta)$ is an abuse of notation for $\varphi(\mathbf{D} \mid \theta(\eta, \zeta))$.

Define for all $\eta$,

$$P'(\eta \mid \bar{H}) = \int f(\eta, \zeta)\, d\zeta, \tag{2.18}$$

and define for all $\eta$, $\zeta$, $P'(\eta, \zeta \mid \bar{H}, \mathbf{D}) = \varphi(\mathbf{D} \mid \eta, \zeta) \cdot f(\eta, \zeta) / \Phi(\mathbf{D} \mid \bar{H})$, motivating the quite natural definition for all $\eta$,

$$P'(\eta \mid \bar{H}, \mathbf{D}) = \int \varphi(\mathbf{D} \mid \eta, \zeta) \cdot f(\eta, \zeta)\, d\zeta / \Phi(\mathbf{D} \mid \bar{H}). \tag{2.19}$$

THEOREM (Savage's Density Ratio). *If*

$$g(\zeta) = f(\eta_0, \xi) / \int f(\eta_0, )\tilde{\zeta}\, d\tilde{\zeta}, \tag{2.20}$$

then

$$L = P'(\eta_0 \mid \bar{H}, \mathbf{D}) / P'(\eta_0 \mid \bar{H}). \tag{2.21}$$

PROOF. Use (2.20) for $g$ in the numerator $\Phi(\mathbf{D} \mid H)$ (2.16) of $L$ (2.4).

Equation (2.21) was given in special approximate forms by Jeffreys (1948), Lindley (1961), and Savage (1959), (1961). Since the first submission of this paper, the authors have found an unpublished general exact statement of (2.21) by Savage (1963). See also Dickey (1968) and Patil (1964).

Formula (2.21) is presented here as a convenient alternative to (2.4) when $f$ belongs to a family of prior distributions conjugate to $\varphi$ in the sense of Raiffa and Schlaifer (1961). For, then, the parameters of $f$ merely change in a simple way. Examples utilizing the most important conjugate family, the beta (Dirichlet), for binomial (multinomial) data, are given below. Examples for normally distributed data will be given elsewhere, including a Bayesian replacement for the usual $F$ test.

The authors have no interest in magical unique prior distributions. Conjugate families are viewed as broad sources of approximate expressions for actual opinions. If several extreme values of prior parameters, chosen to closely bound one's actual opinion, do not lead to weighted likelihood ratios of a single implication, then the data can only be inconclusive to one. The authors realize this may be disturbing to mathematically helpless experimenters who need to "prove something with statistics." Work is needed badly on problems of describing data Bayesianly.

The ratio $L$ of posterior to prior odds may serve as an adequate summary for some scientific experiments. Consider, though, the more general setting of a decision $d$ of $d_H$ for $H$ or $d_{\bar{H}}$ for $\bar{H}$ and a utility function $U(d; \eta, \zeta)$ satisfying

$$\begin{aligned} U(d; \eta, \zeta) &= U(d_H; \eta_0, \zeta) \\ &= U(d_{\bar{H}}; \eta, \zeta) && \eta \neq \eta_0 \\ &= 0 && \text{otherwise.} \end{aligned} \tag{2.22}$$

An optimum decision $d$ is one which maximizes the posterior expected utility, $E[U(d;\eta,\zeta)\,|\,\mathbf{D}]$. As pointed out by Lindley (1961), the utility function (2.22) is the most general one in the testing situation, since the optimum decision $d$ is unaffected by subtracting a function of $\eta$ and $\zeta$ from $U$. We have subtracted $U(d_{\bar{H}};\eta_0,\zeta)$ when $\eta = \eta_0$ and $U(d_H;\eta,\zeta)$ when $\eta \neq \eta_0$.

The posterior expected utility satisfies

$$\begin{aligned} E[U(d; \eta, \zeta)\,|\,D] &= E[U(d_H; \eta_0, \zeta)\,|\,\mathbf{D}, H] \cdot P[H\,|\,\mathbf{D}], \quad \text{for} \quad d = d_H \\ &= E[U(d_{\bar{H}}; \eta, \zeta)\,|\,\mathbf{D}, \bar{H}] \cdot P[\bar{H}, \mathbf{D}], \quad \text{for} \quad d = d_{\bar{H}}. \end{aligned} \tag{2.23}$$

Define the posterior weighted utility ratio $R$ by

$$R = E[U(d_H; \eta_0, \zeta)\,|\,\mathbf{D}, H]/E[U(d_{\bar{H}}; \eta, \zeta)\,|\,\mathbf{D}, \bar{H}]. \tag{2.24}$$

Then the optimal decision is $d_H$ or $d_{\bar{H}}$ according to whether the product

$$O(H) \cdot L \cdot R \tag{2.25}$$

is greater or less than unity (given that the denominator of $R$ is positive). In the following examples, $R$ will be easy to calculate when $U(d_H;\eta_0,\zeta)$ and $U(d_{\bar{H}};\eta,\zeta)$ are polynomials in the coordinates of $\eta$ and $\zeta$.

## 3. Chances.

**3.1.** *A special value for a Bernoulli probability.* For a simple illustration of the general theory without nuisance parameters $\zeta$, let $n$ denote the number of successes

in a sequence $\mathbf{D}$ of $N$ independent Bernoulli trials of unknown success probability $\pi = \theta$. Then

$$\varphi(\mathbf{D} \mid \theta) = \pi^n(1-\pi)^{N-n}. \tag{3.1}$$

To test whether $\pi$ takes on a pre-chosen value $\pi_0$ we let $\eta = \pi$, $\eta_0 = \pi_0$, and test the hypothesis $H: \eta = \eta_0$.

We consider the class of beta prior distributions for $\pi$ under $\overline{H}$. Suppose $\pi \mid \overline{H}$ has a beta distribution with parameters $a, b > 0$; namely, with density on $[0, 1]$, $\pi^{a-1}(1-\pi)^{b-1}/B(a, b)$, where $B(a, b) = \Gamma(a)\,\Gamma(b)/\Gamma(a+b)$. We use the notation

$$\pi \mid \overline{H} \sim \beta(a, b). \tag{3.2}$$

By an application of Bayes' theorem,

$$\pi \mid \overline{H}, \mathbf{D} \sim \beta(\tilde{n}, \tilde{N}-\tilde{n}), \quad \text{where} \tag{3.3}$$

$$\tilde{n} = a+n, \qquad \tilde{N} = a+b+N. \tag{3.4}$$

Hence, by (2.21),

$$L = \pi_0^n(1-\pi_0)^{N-n}B(a, b)/B(\tilde{n}, \tilde{N}-\tilde{n}). \tag{3.5}$$

Jeffreys (1961, page 256), Edwards, Lindman, and Savage (1963, p. 222), and Good (1965, page 41) obtained (3.5) from (2.4).

Recall the classical notation, if $a > 0$, $b \geqq 0$,

$$\begin{aligned}(a, b) &= \Gamma(a+b)/\Gamma(a) \\ &= a(a+1)\cdots(a+b-1), \qquad \text{integer} \quad b > 0.\end{aligned} \tag{3.6}$$

analogous to $a^b$ or $(a+b)^b$, and for which

$$B(a+n, b+m)/B(a, b) = (a, n)(b, m)/(a+b, n+m).$$

Then

$$L = \pi_0^n(1-\pi_0)^{N-n}/[(a, n)(b, N-n)/(a+b, N)]. \tag{3.7}$$

We note that for large, $a$, $b$, and fixed $a/(a+b) = \pi_1$,

$$L \doteq \pi_0^n(1-\pi_0)^{N-n}/[\pi_1^n(1-\pi_1)^{N-n}],$$

the usual likelihood ratio with the simple alternative $\overline{H}: \pi = \pi_1$.

For small $a > 0$, $\Gamma(a) = o(1/a)$, hence $B(a, b) \to \infty$ as $a \to 0+$ or $b \to 0+$ or both. Consequently, $L \to \infty$ for $\mathbf{D}$ containing at least one each of a success and a failure, as the nonintegrable prior, $a = 0$, $b = 0$, is approached. This pseudoprior, expressing "ignorance" to some Bayesians, is analogous in this implication for $L$ to an unbounded-support "uniform" prior for a normal mean. Such infinite posterior odds for $H$ in cases of "ignorance" under $\overline{H}$ received much discussion by Cornfield (1966), whose equation (7.8) violates our assumption (2.20) relating $g$ to $f$.

An interesting approximate form of $L$ for large $\tilde{N}$ and $\tilde{n}$ follows from an application of Stirling's approximation to the complete beta integral,

$$B(\tilde{n}, \tilde{N}-\tilde{n}) \doteq -\tfrac{1}{2}\log\tilde{N} - \tilde{N}I(\hat{\pi}) - \tfrac{1}{2}\log[\hat{\pi}(1-\hat{\pi})] + \tfrac{1}{2}\log(2\pi), \tag{3.8}$$

where the "information" $I(\hat{\pi})$ is given by

$$I(\hat{\pi}) = -\hat{\pi}\log\hat{\pi} - (1-\hat{\pi})\log(1-\hat{\pi}) \tag{3.9}$$

and

$$\hat{\pi} = \tilde{n}/\tilde{N}. \tag{3.10}$$

The decision factor $R$ (2.24) can be calculated from the distribution (3.3). For example, if $U(d_H; \pi_0) = c_1$ and $U(d_{\bar{H}}; \pi \neq \pi_0) = c_2 + c_3\eta^2$, then

$$R = c_1/[c_2 + c_3(\hat{\pi}-\pi_0)^2 + c_3\,\hat{\pi}(1-\hat{\pi})/(\tilde{N}+1)].$$

**3.2.** *Equality of two Bernoulli (binomial) probabilities.* Suppose $n_1$ and $n_2$ are the number of successes in $N_1$ and $N_2$ ($N_1 + N_2 = N$) independent Bernoulli trials with unknown probabilities of success $\pi_1$ and $\pi_2$. Then for the sequences $\mathbf{D}$ of observations and $\boldsymbol{\theta} = (\pi_1, \pi_2)$

$$\varphi(\mathbf{D} \mid \boldsymbol{\theta}) = \textstyle\prod_{i=1}^{2} \pi_i^{n_i}(1-\pi_i)^{N_i - n_i}. \tag{3.11}$$

Define $\eta$ and $\zeta$ by

$$\eta = \pi_1 - \pi_2, \qquad \zeta = \tfrac{1}{2}(\pi_1 + \pi_2). \tag{3.12}$$

We seek to test the natural hypothesis that $\pi_1 = \pi_2$, or with $\eta_0 = 0$, $H: \eta = 0$.

We take independent beta prior distributions for the $\pi_i$'s under $\bar{H}$. Symbolically, for $a_i, b_i > 0$,

$$\pi_i \mid \bar{H} \sim \beta(a_i, b_i). \tag{3.13}$$

By an application of Bayes' theorem, we have

$$\pi_i \mid \bar{H}, \mathbf{D} \sim \beta(\tilde{n}_i, \tilde{N}_i - \tilde{n}_i), \tag{3.14}$$

where

$$\tilde{n}_i = a_i + n_i, \qquad \tilde{N}_i = a_i + b_i + N_i \tag{3.15}$$

with $\pi_1 \mid \bar{H}, \mathbf{D}$ and $\pi_2 \mid \bar{H}, \mathbf{D}$ independent. Thus, to use equation (2.21), we merely need the density at zero of the difference of two independent beta random variables. Direct calculation yields

$$P'(\eta = 0 \mid \bar{H}) = \frac{B(a_1 + a_2 - 1, b_1 + b_2 - 1)}{B(a_1, b_1)B(a_2, b_2)} \tag{3.16}$$

and

$$P'(\eta = 0 \mid D, \bar{H}) = \frac{B(\tilde{n}_1 + \tilde{n}_2 - 1, \tilde{N}_1 + \tilde{N}_2 - \tilde{n}_1 - \tilde{n}_2 - 1)}{B(\tilde{n}_1, \tilde{N}_1 - \tilde{n}_1)B(\tilde{n}_2, \tilde{N}_2 - \tilde{n}_2)}. \tag{3.17}$$

$L$ is then the ratio (3.17)/(3.16),

$$L = \frac{(a_1 + a_2 - 1, n_1 + n_2)(b_1 + b_2 - 1, N_1 - n_1 + N_2 - n_2)}{(a_1 + a_2 + b_1 + b_2 - 2, N_1 + N_2)} \cdot \left[\frac{(a_1, n_1)(b_1, N_1 - n_1)}{(a_1 + b_1, N_1)}\right]^{-1} \left[\frac{(a_2, n_2)(b_2, N_2 - n_2)}{(a_2 + b_2, N_2)}\right]^{-1}. \tag{3.18}$$

Jeffreys (1948, page 263) used essentially (2.4) to derive an $L$, but with numerator and denominator $\Phi$ conditioned on the margin $n_1 + n_2$.

Cases where $a_1+a_2 \leqq 1$ or $b_1+b_2 \leqq 1$ lead to $L=0$, since the denominator (3.16) is then infinite while the numerator (3.17) will probably be finite. In such instances data exhibiting both successes and failures prove $\bar{H}$. A feeling for this pathology can be gained by examining the numerator $\Phi(\mathbf{D} \mid H)$ (2.16) of $L$ (2.4). The distribution of the data under $H$ is a mixture over $\zeta = \pi$, which has the following limiting distribution under $H$. For $r = (a_1+a_2-1)/(b_1+b_2-1)$,

$$\begin{aligned} \zeta &= 0 \quad \text{with probability} \quad 1/(1+r) \\ &= 1 \quad \text{with probability} \quad r/(1+r) \end{aligned} \tag{3.19}$$

as $a_1+a_2 \to 1+$. Hence, under $H$ the data must consist of all successes or all failures, according to the value of $\zeta$. (This result contrasts with $L=\infty$ for the normal-theory Behrens-Fisher hypothesis under the "ignorance" alternative.)

Posterior moments of $\eta$ and $\zeta$ under $H$ and $\bar{H}$, useful in calculating the factor $R$ (2.24) for the decision criterion (2.25), follow from the distributions

$$\zeta \mid H, \mathbf{D} \sim \beta(\tilde{n}_1+\tilde{n}_2-1, \tilde{N}_1+\tilde{N}_2-\tilde{n}_1-\tilde{n}_2-1) \tag{3.20}$$

and (3.14) for $\pi_1, \pi_2 \mid \bar{H}$, $\mathbf{D}$. For example, if

$$\begin{aligned} U(d; \eta, \zeta) &= c_1+c_2\zeta && \text{if} \quad d=d_H \quad \text{and} \quad \eta=0, \\ &= c_3+c_4\zeta+c_5\eta^2 && \text{if} \quad d=d_H \quad \text{and} \quad \eta \neq 0, \\ &= 0 && \text{otherwise.} \end{aligned} \tag{3.21}$$

then

$$\begin{aligned} R = [(c_1+c_2(\textstyle\sum a_i-1))/(\sum a_i+\sum b_i-2)] \\ \div [c_3+\tfrac{1}{2}c_4\textstyle\sum \tilde{n}_i/\tilde{N}_i+c_5\sum(\tilde{n}_i^2+\tilde{n}_i)/(\tilde{N}_i^2+\tilde{N}_i)-2c_5\prod \tilde{n}_i/\tilde{N}_i]. \end{aligned} \tag{3.22}$$

**3.3.** *An invariance property.* Several persons, including the referee, have asked the authors whether $L$ is invariant with respect to the choice of defining transformation $\boldsymbol{\eta}(\boldsymbol{\theta})$ for $H$. The answer is yes for an integrable prior density $f$, if $f$ and $g$ induce the densities $f^*$ and $g^*$ for another parametrization $\boldsymbol{\eta}^*(\boldsymbol{\eta}, \boldsymbol{\zeta})$, $\boldsymbol{\zeta}^*(\boldsymbol{\eta}, \boldsymbol{\zeta})$.

$$\boldsymbol{\eta}^*(\boldsymbol{\eta}_0, \boldsymbol{\zeta}) = \boldsymbol{\eta}^*(\boldsymbol{\eta}_0) = \boldsymbol{\eta}_0{}^* \tag{3.23a}$$

$$\boldsymbol{\eta}(\boldsymbol{\eta}_0{}^*, \boldsymbol{\zeta}^*) = \boldsymbol{\eta}(\boldsymbol{\eta}_0{}^*) = \boldsymbol{\eta}_0, \tag{3.23b}$$

$$f^*(\boldsymbol{\eta}^*, \boldsymbol{\zeta}^*) = f(\boldsymbol{\eta}, \boldsymbol{\zeta}) \cdot J(\partial\boldsymbol{\xi}/\partial\boldsymbol{\xi}^*) \tag{3.24}$$

$$g^*(\boldsymbol{\zeta}^*) = g(\boldsymbol{\zeta}) \cdot J(\partial\boldsymbol{\zeta}/\partial\boldsymbol{\zeta}^*)_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}. \tag{3.25}$$

This answer is obvious in regard to the expression (2.4) for $L$ as a ratio of prior moments $\Phi$.

To prove

$$L = P^{*\prime}(\boldsymbol{\eta}_0{}^* \mid \bar{H}, \mathbf{D})/P^{*\prime}(\boldsymbol{\eta}_0{}^* \mid \bar{H}), \tag{3.26}$$

we need merely show that

$$g^*(\boldsymbol{\zeta}^*) = f^*(\boldsymbol{\eta}_0{}^*, \boldsymbol{\zeta}^*)/\textstyle\int f^*(\boldsymbol{\eta}_0{}^*, \tilde{\boldsymbol{\zeta}}^*)\, d\tilde{\boldsymbol{\zeta}}^*. \tag{3.27}$$

But by (3.23b),

$$(\partial/\partial\zeta_j^*)\eta_i(\boldsymbol{\eta}_0^*, \zeta^*) = 0, \qquad \text{each } i, j, \quad \text{so} \tag{3.28}$$

$$J(\partial\xi/\partial\xi^*)_{\eta=\eta_0} = J(\partial\zeta/\partial\zeta^*)_{\eta=\eta_0} \cdot J(\partial\boldsymbol{\eta}/\partial\boldsymbol{\eta}^*)_{\eta=\eta_0}, \tag{3.29}$$

Equations (3.24), (3.29), and (3.25) imply

$$\begin{aligned} &f^*(\boldsymbol{\eta}_0^*, \zeta^*)/\textstyle\int f^*(\boldsymbol{\eta}_0^*, \tilde{\zeta}^*)\, d\tilde{\zeta}^* = \\ &\qquad f(\boldsymbol{\eta}_0, \zeta) \cdot J(\partial\zeta/\partial\zeta^*)_{\eta=\eta_0} / \textstyle\int f(\boldsymbol{\eta}_0, \zeta)\, d\zeta = \\ &\qquad\qquad g(\zeta) \cdot J(\partial\zeta/\partial\zeta^*)_{\eta=\eta_0} = g^*(\zeta^*). \end{aligned} \tag{3.30}$$

The property (3.28) is easily verified for $\eta = \theta_1 - \theta_2$, $\zeta = \frac{1}{2}(\theta_1 + \theta_2)$, $\eta^* = \log(\theta_1/\theta_2)$, $\zeta^* = \log(\theta_1, \theta_2)$, $\eta_0 = \eta_0^* = 0$.

## 4. Multivariate chances.

**4.1.** *A special value for a multinomial mean vector.* Let $\pi(k)$ be a probability mass function on $k = 1, \cdots, K$, and let $\boldsymbol{\theta} = \boldsymbol{\pi}, r = K$. The $N$ independent realizations from $\boldsymbol{\pi}$ with cell frequencies $n(k)$, $\sum n(k) = N$, occur with probability,

$$\varphi(\mathbf{D} \mid \boldsymbol{\theta}) = \textstyle\prod \pi(k)^{n(k)}. \tag{4.1}$$

With $\boldsymbol{\eta} = \boldsymbol{\theta} = \boldsymbol{\pi}$, and $\boldsymbol{\eta}_0 = \boldsymbol{\pi}_0$, consider the hypothesis $H: \boldsymbol{\eta} = \boldsymbol{\eta}_0$, or $\boldsymbol{\pi} = \boldsymbol{\pi}_0$.

We take a Dirichlet prior for $\boldsymbol{\pi}$ under $\overline{H}$, with density on $\sigma_K = \{\boldsymbol{\pi}: \sum_\kappa \pi(k) = 1$ and each $\pi(k) > 0\}$, $\prod_\kappa \pi(k)^{a(k)-1}/B(\mathbf{a})$, each $a(k) > 0$, and $B(\mathbf{a})$ denoting Dirichlet's complete integral,

$$B(\mathbf{a}) = \textstyle\prod \Gamma(a(k))/\Gamma(\sum a(k)). \qquad \text{Symbolically,} \tag{4.2}$$

$$\boldsymbol{\pi} \mid \overline{H} \sim \beta(\mathbf{a}). \tag{4.3}$$

The multinomial analogues of equations (3.3) through (3.5) are

$$\boldsymbol{\pi} \mid \overline{H}, \mathbf{D} \sim \beta(\tilde{\mathbf{n}}), \tag{4.4}$$

$$\tilde{\mathbf{n}} = \mathbf{a} + \mathbf{n}, \tag{4.5}$$

$$L = [\textstyle\prod \pi_0(k)^{n(k)}] B(\mathbf{a})/B(\tilde{\mathbf{n}}). \tag{4.6}$$

Note that

$$B(\mathbf{a}+\mathbf{b})/B(\mathbf{a}) = \textstyle\prod (a(k), b(k))/(\sum a(k), \sum b(k)), \tag{4.7}$$

in the notation (3.6). Hence

$$L = [\textstyle\prod \pi_0(k)^{n(k)}]/[\prod (a(k), n(k))/(\sum a(k), N)]. \tag{4.8}$$

Good (1965, pages 37 and 41) derived (4.6) from (2.4). Good (1965) and (1967) takes special interest in the symmetric-Dirichlet alternative, parameters $a(k) = a$, the "flattening constant"; and the hypothesis $H$ of equal probabilities, namely, with special value $\pi_0(k) \equiv K^{-1}$; then $L = K^{-N}[\Gamma(a)^K/\Gamma(Ka)]/[\prod_\kappa \Gamma(a+n_k)/\Gamma(Ka+N)]$.

Good suggests, among others, the choice of $\mathbf{a}$ to maximize $\Phi(\mathbf{D} \mid \bar{H}) = K^{-N}/L$, and also symmetric priors obtained as mixtures over $a$.

**4.2.** *Equality of several multinomial mean vectors.* The results of Section 3.2 are extended in two directions simultaneously in this section. The first is to test the equality of two multinomial mean vectors. The second is to test the equality of several such vectors.

For each $I = 1, \cdots, I$ let $\pi_i(k)$ be a probability mass function on $k = 1, \cdots, K$ and let $\boldsymbol{\theta} = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_I)$, $r = IK$. The $N$ independent realizations $\mathbf{D}, N_i$ from $\pi_i$, $\sum N_i = N$, with cell frequencies $n_i(k)$, $\sum_k n_i(k) = N_i$, occur with probability

$$\varphi(\mathbf{D} \mid \boldsymbol{\theta}) = \prod\prod\prod \pi_i(k)^{n_i(k)}. \tag{4.9}$$

Define the vectors $\boldsymbol{\eta}_i$ and $\boldsymbol{\zeta}$ by

$$\boldsymbol{\eta}_i = \boldsymbol{\pi}_i - \boldsymbol{\pi}_{i-1}, \qquad i = 2, \cdots, I \quad \text{and} \tag{4.10}$$

$$\boldsymbol{\zeta} = I^{-1} \sum \boldsymbol{\pi}_i. \tag{4.11}$$

Each $\boldsymbol{\eta}_i$ and $\boldsymbol{\zeta}$ is $K$ dimensional but is subject to the constraints $\sum_k \eta_i(k) = 0$ and $\sum_k \zeta(k) = 1$. We test the hypothesis of identical probability distributions $\boldsymbol{\pi}_i$,

$$H: \boldsymbol{\eta}_2 = \boldsymbol{\eta}_3 = \cdots = \boldsymbol{\eta}_I = \mathbf{0}, \qquad s = (I-1)K. \tag{4.12}$$

We take independent Dirichlet prior distributions for the $\boldsymbol{\pi}_i$'s under $\bar{H}$. Symbolically,

$$\boldsymbol{\pi}_i \mid \bar{H} \sim \beta(\mathbf{a}_i). \tag{4.13}$$

Then, independently in $i$,

$$\boldsymbol{\pi}_i \mid \bar{H}, \mathbf{D} \sim \beta(\tilde{\mathbf{n}}_i) \qquad \text{where} \tag{4.14}$$

$$\tilde{\mathbf{n}}_i = \mathbf{a}_i + \mathbf{n}_i. \tag{4.15}$$

We calculate $L$ from its original form (2.4) as a ratio of subjective mixtures $\Phi$ of $\varphi(\mathbf{D} \mid \boldsymbol{\theta})$. The joint density of the $\boldsymbol{\pi}_i$'s yields

$$\boldsymbol{\zeta} \mid H \sim \beta(\textstyle\sum \mathbf{a}_i - (I-1)\mathbf{1}), \qquad \text{where} \tag{4.16}$$

$$\mathbf{1} = (1, \cdots, 1)', \tag{4.17}$$

and so for the numerator of $L$ (2.4) we have

$$\begin{aligned} \Phi(\mathbf{D} \mid H) &= E_{\zeta \mid H}\, \varphi(\mathbf{D} \mid \boldsymbol{\theta}) \\ &= B(\textstyle\sum \tilde{\mathbf{n}}_i - (I-1)\mathbf{1}) / B(\sum \mathbf{a}_i - (I-1)\mathbf{1}). \end{aligned} \tag{4.18}$$

The denominator is given by

$$\Phi(\mathbf{D} \mid \bar{H}) = E_{\theta \mid H}\, \varphi(\mathbf{D} \mid \boldsymbol{\theta}) = \prod [B(\tilde{\mathbf{n}}_i)/B(\mathbf{a}_i)]. \tag{4.19}$$

By (4.7),

$$L = \frac{\prod_\kappa(\sum_i a_i(k) - I + 1, \sum_i n_i(k))}{(\sum_{i,\kappa} a_i(k) - IK + K, \sum_{i,\kappa} n_i(k))} \div \prod_i \frac{\prod_\kappa(a_i(k), n_i(k))}{(\sum_\kappa a_i(k), \sum_\kappa n_i(k))} \tag{4.20}$$

For identical symmetric-Dirichlet prior distributions, $a_i(k) \equiv a$,

$$L = \frac{\prod_\kappa(I(a-1)+1, \sum_i n_i(k))}{(IK(a-1)+K, \sum_{i,\kappa} n_i(k))} \div \prod_i \frac{\prod_\kappa(a, n_i(k))}{(Ka, \sum_\kappa n_i(k))} \tag{4.21}$$

a special case of Good's (1965, equation (6.2)) statistic for the hypothesis of independence in a two-way contingency table.

The multivariate analogue of (3.20) is

$$\zeta \mid H, \mathbf{D} \sim \beta(\sum \tilde{\mathbf{n}}_i - (I-1)\mathbf{1}). \tag{4.22}$$

**5. The order of a finite-state Markov chain.** Let $\mathbf{D}$ be a length-$N$ realization of a $K$-state Markov chain known to be of order at most $\bar{h} > 0$,

$$\mathbf{D} = (k_1, k_2, \cdots, k_N)', \qquad \text{each} \quad k_n \in \{1, \cdots, K\} \tag{5.1}$$

$$\varphi(\mathbf{D} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} \pi_{k_{n-\bar{h}}, k_{n-\bar{h}+1}, \cdots, k_{n-1}}(k_n). \tag{5.2}$$

Regretably, we take the $\bar{h}$ initial states $k_{1-\bar{h}}, k_{2-\bar{h}}, \cdots, k_0$ as a non-informative given condition to all the probabilities. (It has been said that for the subjectivist, all probabilities are conditional.) In practice, this restriction will usually amount to ignoring information about $\boldsymbol{\theta}$ imparted by just the first $\bar{h}$ states, a triviality if $\bar{h}/N$ is small and zero transition probabilities are not a consideration.

Suppose one suspects the order of the chain to be $h$, $0 \leqq h < \bar{h}$, $h$ fixed. Denote by $\mathbf{i}$ an $(\bar{h}-h)$-tuple of states and by $\mathbf{j}$ an $h$-tuple of states ($\mathbf{j}$ absent if $h = 0$). Then redenote the transition probabilities,

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_j)_{\text{all } j}, \qquad \boldsymbol{\theta}_j = (\pi_{ij})_{\text{all } i}, \qquad \pi_{ij} = (\pi_{ij}(1), \cdots, \pi_{ij}(K))', \tag{5.3}$$

$$\varphi(\mathbf{D} \mid \boldsymbol{\theta}) = \prod_{ijk} \pi_{ij}(k)^{n_{ij}(k)}, \tag{5.4}$$

where the transition counts $n_{ij}(k)$ suffice for $\mathbf{D}$,

$$n_{ij}(k) = \sum_{n=1}^{N} \delta_{ijk}(k_{n-\bar{h}}, k_{n-\bar{h}+1}, \cdots, k_n), \tag{5.5}$$

$$\mathbf{n}_{ij} = (n_{ij}(1), \cdots, n_{ij}(K))'. \tag{5.6}$$

If $H_j$ denotes the hypothesis that the probability $\pi_{ij}(k)$ of transition from the

sequence $\mathbf{j}$ of $h$ states to each state $k$ does not depend on the sequence $\mathbf{i}$ of $\bar{h}-h$ states immediately preceding the states $\mathbf{j}$,

$$H_j: \pi_{ij} = \pi_{i^*j}, \qquad \text{all } \mathbf{i}, \mathbf{i}^*, \text{ then} \tag{5.7}$$

$$H = \cap_{\text{all } j} H_j. \tag{5.8}$$

The similarity of each hypothesis $H_j$ to the hypothesis treated in the previous section, equality of multinomial mean vectors, is more than just notational. Denote by $\mathbf{D}_{ij}^*$ data from a hypothetical fixed-length ($\sum_k n_{ij}(k)$) sequence of multinomial trials having the vector of cell counts $\mathbf{n}_{ij}$ with vector of means $\pi_{ij}$. Then with $\mathbf{D}_j{}^* = (\mathbf{D}_{ij}^*)_{\text{all } i}$,

$$\varphi(\mathbf{D} \mid \theta) = \prod_j \varphi_j{}^*(\mathbf{D}_j{}^* \mid \theta_j), \tag{5.9}$$

where, as in (4.9),

$$\varphi_j{}^*(\mathbf{D}_j{}^* \mid \theta_j) = \prod_{ik} \pi_{ij}(k)^{n_{ij}(k)}. \tag{5.10}$$

The $\mathbf{D}_{ij}^*$'s are independent given $\theta$.

We invoke the likelihood principle (Lindley (1965), for example) to make inference about $\theta_j$ from $\mathbf{D}_j{}^*$. For a fixed $\mathbf{j}$, given $\bar{H}_j$, we take the $K^{\bar{h}-h}$ many $\pi_{ij}$'s to have prior independence and, as in (4.13)

$$\pi_{ij} \mid \bar{H}_j \sim \beta(\mathbf{a}_{ij}). \tag{5.11}$$

(Martin (1967) has used prior Dirichlet distributions for Markov-chain control-theory problems.)

Then, as in (4.14),

$$\pi_i \mid \bar{H}_j, \mathbf{D}_j{}^* \sim \beta(\mathbf{a}_{ij} + \mathbf{n}_{ij}) \tag{5.12}$$

Hence $L_{D_j{}^*}(H_j) = [P(H_j \mid \mathbf{D}_j{}^*)/P(\bar{H}_j \mid \mathbf{D}_j{}^*)]/[P(H_j)/P(\bar{H}_j)]$ is given by equation (4.20) with $i$ replaced $\mathbf{i}$, $I$ replaced by $K^{\bar{h}-h}$, $a_i(k)$ replaced by $a_{ij}(k)$, and $n_i(k)$ replaced by $n_{ij}(k)$. If prior opinion is invariant in the index $\mathbf{ij}k$, a similar statement holds for equation (4.21).

For example, to test whether a first-order Markov chain is an independent sequence, $\mathbf{ij}k = i, k$, say $a_{ij}(k) \equiv a$, $L_D(H) = L_{D_j}^*(H_j) = L$, as given by (4.21).

Now, on the one hand, if

$$P(H) = P(H_j), \quad \text{each } \mathbf{j}, \tag{5.13}$$

and if given $A$ the $\theta_j$'s are prior independent for $A = \bar{H}$ and hence for $A = H$, then $\Phi(\mathbf{D} \mid A) = \prod_j \Phi_j{}^*(\mathbf{D}_j{}^* \mid A)$ for $A = \bar{H}$ and for $A = H$, and so

$$L_D(H) = \prod_j L_{D_j{}^*}(H_j). \tag{5.14}$$

On the other hand, if the $\theta_j$'s are prior independent, hence

$$P(H) = \prod_j P(H_j). \tag{5.15}$$

then the $\theta_j$'s are also posterior independent,

$$P(H \mid \mathbf{D}) = \prod_j P(H_j \mid \mathbf{D}_j^*) \tag{5.16}$$
$$= 1/\prod_j [1 + L_{D_j^*}(H_j)^{-1} P(\bar{H}_j)/P(H_j)].$$

Equations (5.15) and (5.16) yield

$$L_D(H) = \{\prod_j [1 + P(\bar{H}_j)/P(H_j)] - 1\}/ \tag{5.17}$$
$$\{\prod_j [1 + L_{D_j^*}(H_j)^{-1} P(\bar{H}_j)/P(H_j)] - 1\}.$$

Methods will be published elsewhere for more general prior distributions for intersecting hypotheses $H_j$.

**Acknowledgments.** The authors are grateful to a referee for requesting the statement of equation (2.21) in theorem form.

## REFERENCES

[1] ANDO, ALBERT and KAUFMAN, G. M. (1965). Bayesian analysis of the independent multinormal process—neither mean nor precision known. *J. Amer. Statist. Assoc.* **60** 347–358.
[2] BARNARD, G. A. (1964). Unpublished series of lectures on statistical inference. Given at the National Institutes of Health.
[3] CORNFIELD, JEROME (1966). A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *J. Amer. Statist. Assoc.* **61** 577–594.
[4] DICKEY, JAMES M. (1967a). A Bayesian hypothesis-decision procedure. *Ann. Inst. Statist. Math.* **19** 367–369.
[5] DICKEY, JAMES M. (1967b). Letter to the editor. *The American Statistician* **21** No. 4, 49.
[6] DICKEY, JAMES M. (1968). Three multidimensional-integral identities with Bayesian applications. *Ann. Math. Statist.* **39** 1615–1627.
[7] EDWARDS, WARD, LINDMAN, HAROLD, and SAVAGE, LEONARD J. (1963). Bayesian statistical inference for psychological research. *Psychological Rev.* **70** 193–242. ((1965). Reprinted in *Readings In Mathematical Psychology* **2** ed. Luce, R. Duncan, Bush, Robert R., and Galanter, Eugene. Wiley, 519–568).
[8a] GOOD, I. J. (1950). *Probability and the Weighing of Evidence.* Hafner, New York.
[8] GOOD, I. J. (1965). *The Estimation of Probabilities.* M.I.T. Press, Cambridge.
[9] GOOD, I. J. (1966). How to estimate probabilities. *J. Inst. Math. Appl.* **2** 364–383.
[10] GOOD, I. J. (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. Ser. B.* **29** 399–431.
[11] JEFFREYS, HAROLD (1961). *Theory of Probability* 3rd ed. Clarendon Press, Oxford.
[12] LINDLEY, D. V. (1961). The use of prior probability distributions in statistical inference and decision. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* Univ. of Calif. Press, 453–468.
[13] LINDLEY, D. V. (1965). *Introduction to Probability and Statistics From A Bayesian Viewpoint: Part-Two Inference.* Cambridge Univ. Press, Cambridge.
[14] LINDLEY, D. V. (1968). The choice of variables in multiple regression. *J. Roy. Statist. Soc. Ser. B.* **30** 31–66.
[15] MARTIN, J. J. (1967). *Bayesian Decision Problems and Markov Chains.* Wiley, New York.
[16] PATIL, V. H. (1964). The Behrens-Fisher problem and its Bayesian solution. *Journal Indian Statist. Assoc.* **2** 21–31.
[17] RAIFFA, HOWARD, and SCHLAIFER, ROBERT (1961). *Applied Statistical Decision Theory.* Harvard Univ. Press, Boston.
[18] SAVAGE, LEONARD J. (1961). *The Subjective Basis of Statistical Practice.* Unpublished Univ. of Mich. Notes.

[19] SAVAGE, LEONARD J. (1962). *The Foundations of Statistical Inference*, Joint Statistics Seminar at the University of London. Methuen, London.
[20] SAVAGE, LEONARD J. (1963). Notes from Univ. of Michigan course on the Foundations of Statistics.
[21] WALD, A. (1947). *Sequential Analysis*, Wiley, New York.
[22] WISNIEWSKI, T. K. M. (1968). Testing for homogeneity of a binomial series. *Biometrika* **55** 426–428