ON THE FOUNDATIONS OF STATISTICAL INFERENCE: BINARY EXPERIMENTS[1]

By Allan Birnbaum

Institute of Mathematical Sciences, New York University

0. Introduction and summary. In Part A (Sections 1–5) the canonical forms of experiments concerning two simple hypotheses, and their partial ordering, are discussed. It is proved that every such experiment is a mixture (in a probability sense) of simple experiments whose sample spaces contain only two points. In Parts B (Sections 6–8) some general aspects of inference and decision problems are discussed in the usual theoretical framework, in which the overall mathematical model of an experiment is the frame of reference for all interpretations of outcomes.

In Part C (Sections 9–16), attention is directed to that traditional function and basic problem of mathematical statistics, called here "informative inference," whose object is to recognize and report in appropriate objective terms those features of experimental outcomes which constitute statistical evidence relevant to hypotheses (or parameter values) of interest. The mathematical structure of statistical evidence, and its qualitative and quantitative properties, are analyzed by application of (1) the mathematical results of Part A, which show that conditional experimental frames of reference (in the mixture sense) exist and are recognizable much more widely than has previously been realized; and (2) a single extra-mathematical proposition which many statisticians seem inclined to accept as appropriate for purposes of informative inference, a "principle of conditionality" which asserts that any outcome of any experiment which is a mixture of component experiments should be interpreted in the same way as if it were an outcome of just a corresponding component experiment (with the overall mixture structure otherwise ignored). This analysis establishes the likelihood function as the appropriate basis from which statistical inferences can be made directly without other reference to the structure of an experiment. For the numerical values of the likelihood function, this analysis provides direct interpretations in terms of probabilities of errors. These probabilities admit frequency interpretations of the usual kind, but they are not in general defined with reference to the specific experiment from which an outcome is obtained: they express intrinsic objective properties of the likelihood function itself, which this analysis shows to be appropriately relevant and directly useful for purposes of informative inference. The relations of this analysis of problems of informative inference to problems of testing statistical hypotheses, decision-making, conclusions, and Bayesian treatments of inference problems are discussed briefly.

Received July 30, 1960; revised October 11, 1960.

[1] Prepared under the sponsorship of the Office of Naval Research, United States Navy, Contract No. Nonr-285 (38).

Generalizations of these mathematical results and their interpretations for problems involving more than two simple hypotheses will be given in a following paper.

A. MATHEMATICAL DEVELOPMENTS

1. The canonical form of a binary experiment. We consider a given experiment E, assuming that questions of experimental design, including those of choice of a sample size or possibly a sequential sampling rule, have been dealt with, and that the sample space of possible outcomes x of E is a specified set $S = \{x\}$. We assume that each of the possible distributions of X is represented by a specified elementary probability function $f_i(x)$: if the hypothesis H_i is true, the probability that E yields an outcome x in A is

$$P_i(A) = \int_A f_i(x)\, d\mu(x), \tag{1.1}$$

where μ is a specified σ-finite measure on S, and A is any measurable set. We assume until otherwise stated that there are only two possible distributions, so that $i = 1$ or 2. Such experiments will be termed *binary experiments.*

Discussions of statistical inference problems concerning binary experiments usually specify at the outset that the problem under consideration is that of testing the simple hypothesis H_1 against the simple alternative H_2, or that of making one of two specified decisions, on the basis of an observed value of X. These discussions seem to assume tacitly that such formulations are the only ones of possible interest, or at least the only ones sufficiently definite to allow satisfactory theoretical treatment and objective practical application. (We do not consider here formulations in which it is assumed that there exist probabilities of the hypotheses themselves, Prob (H_i), $i = 1, 2$, in some sense.) We begin however with a less formal but broader specification: the general goal is to make inferences from an observed value of X to the hypotheses. Our purpose is to show that this broader specification suffices to guide a useful analysis of the mathematical structure of any given experiment E, an analysis which exhibits some new mathematical properties of experiments that are of intrinsic interest and relevance for statistical inference in general, and throws some new light on more specialized formulations of inference problems.

For any given binary experiment E, let

$$r = r(x) = \log [f_2(x)/f_1(x)]. \tag{1.2}$$

It is well known that r is a (minimal) sufficient statistic. Let

$$F_i(r) = \text{Prob}\,[r(X) \leqq r \mid H_i], \qquad i = 1, 2. \tag{1.3}$$

In general $r(X)$ is a generalized random variable in the sense that it may assume infinite values with positive probability under one or both hypotheses; correspondingly, in general F_1 and F_2 are generalized cumulative distribution func-

tions (c.d.f's.). The pair of distributions F_1, F_2 of r may be taken as a canonical form of any binary experiment E.

A canonical form which is more convenient for many purposes is obtained as follows: Let

$$u(r, z) = zF_1(r) + (1 - z)F_1(r-), \tag{1.4}$$

for $0 < z \leqq 1$ and $-\infty \leqq r \leqq \infty$. If Z is an auxiliary randomization variable, that is, a random variable having under each hypothesis the same uniform distribution on the unit interval, $0 < Z \leqq 1$, independent of X, then $U = u(r(X), Z)$ may be called the continuous probability integral transform of $R = r(X)$, since

$$\text{Prob } (u(R, Z) \leqq u \mid H_1) = u, \qquad \text{for } 0 \leqq u \leqq 1. \tag{1.5}$$

Since r is a function of $u(r, z)$, the latter is a sufficient statistic. For each u, let $v(u) = \text{Prob } [u(R, Z) \leqq u \mid H_2]$, $0 \leqq u \leqq 1$. The function (c.d.f.) $v(u)$ may be regarded as the *canonical form* of the given binary experiment E as was pointed out in [1]. (For each u, by the fundamental lemma of Neyman and Pearson a best test of size $1 - u$ is one which rejects H_1 when $u(r, z)$ exceeds u; with this test, the probability of a Type II error is $v(u)$. The latter is well known to a convex function of u.) Since $v(u)$ is convex, it is continuous, except possibly at $u = 1$, where $v(1) = 1$ always.

Conversely, each convex c.d.f. $v(u)$ on the closed unit interval is the canonical form of some binary experiment. For if $v(u)$ is convex and $v(0) = 0$, $v(1) = 1$, let $f_2(u)$ denote $v'(u)$, the right derivative of $v(u)$, for each $u < 1$, and let $f_2(1) = \infty$. Let $f_1(u) = 1$, $0 \leqq u \leqq 1$. Then the binary experiment E represented by the elementary probability functions $f_1(u)$, $f_2(u)$ (with respect to Lebesgue measure) has the canonical form $v(u)$, as is readily verified.

It is often convenient to consider a binary experiment as represented by the graph of its "$v(u)$ curve," with the latter supplemented by a vertical line-segment if necessary so as to give in all cases a graphically-continuous convex curve from (0, 0) to (1, 1).

2. Simple binary experiments. A binary experiment with $v(u) \equiv u$ is trivial in the sense that its sufficient statistic $r = r(x)$ has the same distribution under each hypothesis. Such experiments will be called *uninformative*, and all other experiments will be called *informative*.

A binary experiment will be called *simple* if its sufficient statistic r assumes at most two distinct values, $r_1 \leqq r_2$, (with exceptions on sets of points x having probability 0 under each hypothesis). A binary experiment which is not simple will be called *composite*. In an informative simple binary experiment, we have $r_1 < r_2$, each value having positive probability under at least one hypothesis. In any such experiment, let

$$p_i = \text{Prob } [r(X) = r_2 \mid H_i], \qquad \text{and} \quad q_i = 1 - p_i, \qquad \text{for } i = 1, 2. \tag{2.1}$$

Then $0 \leqq p_1 < p_2 \leqq 1$, or $0 \leqq q_2 < q_1 \leqq 1$; the point (q_1, q_2) characterizes any such experiment, since its $v(u)$ curve consists of two line segments connecting successively the points $(0, 0)$, (q_1, q_2), $(1, 1)$.

Conversely, every such $v(u)$ curve, or every point (q_1, q_2) with $0 \leqq q_2 < q_1 \leqq 1$, characterizes an informative simple binary experiment. For consider any such pair and the experiment E consisting of a single Bernoulli trial such that

$$q_i = \text{Prob}\,[X = 0 \mid H_i], \quad \text{and} \tag{2.2}$$

$$p_i = 1 - q_i = \text{Prob}\,[X = 1 \mid H_i], \qquad i = 1, 2.$$

Its sufficient statistic is

$$r(x) = \begin{cases} r_1 \equiv \log (q_2/q_1) & \text{if } x = 0, \\ r_2 \equiv \log (p_2/p_1) & \text{if } x = 1. \end{cases} \tag{2.3}$$

Any such experiment may be characterized alternatively by a point (r_1, r_2) satisfying $-\infty \leqq r_1 < 0 < r_2 \leqq \infty$, that is by a point in the second quadrant of the (r_1, r_2)-plane excluding the coordinate axes but including all points with one or both coordinates infinite.

A third representation of any informative simple binary experiment is given by the ordered pair (L_1, L_2) of possible values of the likelihood ratio statistic:

$$L_1 = q_2/q_1 = e^{r_1}, \; L_2 = p_2/p_1 = e^{r_2}, \qquad 0 \leqq L_1 < 1 < L_2 \leqq \infty, \tag{2.4}$$

so that $q_1 = (L_2 - 1)/(L_2 - L_1)$ and $q_2 = L_1 q_1$. A fourth representation is given by considering the only nontrivial nonrandomized best test of H_1 against H_2, which rejects H_1 just when $r(x) = r_2$; the probabilities of errors of Types I and II respectively are $(\alpha, \beta) = (p_1, q_2)$, which satisfy $\alpha + \beta < 1$. A fifth useful representation of any such experiment is by means of a stochastic matrix:

$$E = \begin{pmatrix} q_1 & p_1 \\ q_2 & p_2 \end{pmatrix}. \tag{2.5}$$

An uninformative simple binary experiment is represented by $(r_1, r_2) = (0, 0)$, or by $(L_1, L_2) = (1, 1)$, or by $(q_1, q_2) = (q_1, q_1)$ for any q_1, or by $(\alpha, \beta) = (\alpha, 1 - \alpha)$ for any α.

EXAMPLE 1.

"One toss of a coin" experiments. As indicated above, every simple binary experiment is equivalent to an experiment consisting of a single observation on a Bernoulli random variable X with possible values 0 or 1 only.

EXAMPLE 2.

A Wald *sequential probability ratio test* between two simple hypotheses, in special cases including certain tests on a binomial parameter (the cases in which there is "no excess at termination"), is based on a sequential sampling rule which allows only two values for the likelihood ratio statistic, or for $r(x)$. In many other cases, such tests might be called approximately simple

in the sense that under each hypothesis the probability of $r(X) \doteq r_1$ or r_2 is very near unity.

EXAMPLE 3.

Communication channels. In communication theory (information theory), a communication channel (without memory) is any structure which can receive at one point any one of a specified set of "input signals" and deliver at another point one of a designated set of "output signals", the respective probabilities of the latter depending only upon the selected input signal. In the case of just two input signals, which we may denote by H_1, H_2, we have a *binary channel*; we may denote the set of possible output signals by $S = \{x\}$, and the respective probabilities of subsets A of S by $P_i(A)$, $i = 1, 2$. Thus each such communication channel is mathematically equivalent to a binary experiment, and conversely. If $x = 0$ or 1 only, we have a *simple binary* ("two-by-two") *channel*, equivalent to a simple binary experiment. Here (α, β) describe completely the structure of "noise" in the channel: α is the probability that transmission of H_1 will lead to receiving of $x = 1$, and β is the probability that transmission of H_2 will lead to receiving of $x = 0$.

Noisy channels in series. It is convenient to introduce some techniques required below as an elaboration of the present example. Let channel E have inputs H_1, H_2, outputs $x = 0$ or 1, and noise parameters (α, β). Let channel E' have inputs $x = 0$ or 1, outputs $x' = 0$ or 1, and noise parameters (α', β'). Then the channel E^* consisting of E followed by E' has inputs H_1, H_2, and outputs $x' = 0$ or 1. It is useful to write $E^* = EE'$, since if

$$E = \begin{pmatrix} q_1 & p_1 \\ q_2 & p_2 \end{pmatrix} \quad \text{and} \quad E' = \begin{pmatrix} q_1' & p_1' \\ q_2' & p_2' \end{pmatrix}, \tag{2.6}$$

then

$$\begin{aligned} E^* &= \begin{pmatrix} q_1 & p_1 \\ q_2 & p_2 \end{pmatrix}\begin{pmatrix} q_1' & p_1' \\ q_2' & p_2' \end{pmatrix} = EE' \\ &= \begin{pmatrix} q_1 q_1' + p_1 q_2' & q_1 p_1' + p_1 p_2' \\ q_2 q_1' + p_2 q_2' & q_2 p_1' + p_2 p_2' \end{pmatrix} = \begin{pmatrix} q_1^* & p_1^* \\ q_2^* & p_2^* \end{pmatrix}. \end{aligned} \tag{2.7}$$

The noise parameters of E^* are

$$\begin{aligned} (\alpha^*, \beta^*) &= (p_1^*, q_2^*) \\ &= ((1 - \alpha)\alpha' + \alpha(1 - \beta'), \beta(1 - \alpha') + (1 - \beta)\alpha'). \end{aligned} \tag{2.8}$$

The other representations of E^* include

$$L_1^* \equiv q_2^*/q_1^* = (q_2 q_1' + p_2 q_2')/(q_1 q_1' + p_1 q_2'),$$

and (2.9)

$$L_2^* \equiv p_2^*/p_1^* = (q_2 p_1' + p_2 p_2')/(q_1 p_1' + p_1 p_2').$$

If $q_2' = 0$ but $p_1' > 0$, we may say that E' has noise affecting only the transmitted signal $x = 0$; in this case we may also say that E' has noise which degrades only the received signal $x' = 1$, since the received signal $x' = 0$ is known with certainty to follow from a transmitted signal $x = 0$, while a received signal $x' = 1$ is known to be possible following either transmitted signal $x = 0$ or 1. In such a case we have $L_1^* = q_2/q_1 \equiv L_1$ and

$$L_2^* = (p_2 + p_1'q_2)/(p_1 + p_1'q_1) < p_2/p_1 = L_2 \tag{2.10}$$

(assuming $p_1 < p_2$, the remaining case being trivial). Similarly if $p_1' = 0$ but $q_2' > 0$, E' has noise affecting only $x = 1$ and degrading only $x' = 0$, and $L_2^* = p_2/p_1 \equiv L_2$,

$$L_1^* = (q_2 + q_2'p_2)/(q_1 + q_2'p_1) > q_2/q_1 \equiv L_1 \tag{2.11}$$

(assuming the nontrivial case $p_1 < p_2$). It is easily verified that every channel E' is equivalent to a pair of channels in series, $E = E_1 E_2$, where E_1 has noise affecting at most the signal $x = 1$, and E_2 has noise affecting at most the signal $x = 0$.

It follows that for any simple binary channels E, with parameters (L_1, L_2), and E', the channel $E^* = EE'$ has parameters (L_1^*, L_2^*) satisfying $L_1 \leqq L_1^* \leqq 1 \leqq L_2^* \leqq L_2$. And conversely, if E and E^* are channels with parameters satisfying these inequalities, then there exists a channel E' such that $E^* = EE'$. Since $r_i = \log L_i$, these inequalities may be written

$$r_1 \leqq r_1^* \leqq 0 \leqq r_2^* \leqq r_2 . \tag{2.12}$$

EXAMPLE 4.

Significance Tests. In every binary experiment, if the outcome x is to be reported only by a conclusion of the form "reject H_1" or "accept H_1" based on a specified *significance test* with error-probabilities (α, β), then the over-all procedure is formally a simple binary experiment, with $L_1 = \beta/(1 - \alpha)$, $L_2 = (1 - \beta)/\alpha$.

3. The partial ordering of binary experiments. In the theory of comparison of experiments [2], an experiment E is called *at least as informative* as another experiment E^* if and only if it is possible to use E, possibly supplemented by use of an auxiliary randomization variable, to construct an experiment equivalent to E^*. (We depart from the usual terminology, in which "more informative than" is used so as to include the case of equivalence.)

To denote that E is at least as informative as E^*, we write $E \geqq E^*$ or $E^* \leqq E$. It is also convenient to denote this relation by writing that E *contains* E^*, since this terminology has been used in connection with communication channels [3].

If $E \geqq E^*$ and $E^* \geqq E$, we write $E = E^*$ to denote that E is *equivalent* to E^*. We write $E \neq E^*$ to denote that E and E^* are not equivalent. If $E \geqq E^*$ and $E \neq E^*$, we write $E > E^*$ to denote that E is *more informative than* E^*. If neither $E \geqq E^*$ nor $E^* \geqq E$ holds, E and E^* are *not comparable*.

It is well known that, for binary experiments E: $v(u)$ and E^*: $v^*(u)$, we have $E \geqq E^*$ if and only if $v(u) \leqq v^*(u)$ for $0 \leqq u \leqq 1$. In the case of simple binary experiments E: (r_1, r_2) and E^*: (r_1^*, r_2^*), it is readily verified that this condition specilizes to: $E \geqq E^*$ if and only if $r_1 \leqq r_1^* \leqq r_2^* \leqq r_2$; that is, if and only if the *interval* (r_1, r_2) contains the *interval* (r_1^*, r_2^*).

The partial ordering of simple binary experiments determined by the relation $\geqq$ is conveniently represented graphically in the (r_1, r_2) plane. $E > E^*$ denotes that (r_1^*, r_2^*) is closer than (r_1, r_2) to $(0, 0)$ in the sense that at least one of its coordinates is closer to 0 and neither is farther. In a case of non-comparability, one of the points (r_1, r_2), (r_1^*, r_2^*) lies to the upper-right of the other.

Any finite or infinite set of experiments will be called *strictly ordered* if, of every pair in the set, one is more informative. Each such set of experiments corresponds to a subset of the points (r_1, r_2) of some graphically-continuous nonincreasing curve from $(-\infty, \infty)$ to $(0, 0)$. Any such set of experiments has a paramatric representation $(r_1[d], r_2[d])$, with $r_1[d]$ nondecreasing and $r_2[d]$ nonincreasing in d, where d has a specified range.

4. Mixtures of simple binary experiments. If various experiments are possible for a given inference problem, and if one of these is selected for use by means of a specified random device unrelated to the hypotheses, the over-all procedure is called a *mixture of experiments*, or a *mixture experiment*. Since each simple binary experiment is represented by a point (r_1, r_2) in the range described above, the various (generalized) cumulative distribution functions $G(r_1, r_2)$ on that range correspond to the possible mixtures of simple binary experiments. For any such distribution G, we write E_G to designate the (mixture) experiment consisting of the selection of a simple experiment (r_1, r_2) by use of a random device corresponding to G, and the observation of the outcome of one trial of the selected experiment; the simple experiments will be called *components* of E_G.

Any such mixture experiment E_G has the generic sample point $x = (r_1, r_2, r_3)$, where (r_1, r_2) is the selected simple experiment and r_3 is the observed outcome of that experiment, $r_3 = r_1$ or r_2. To determine the sufficient statistic $r(x) = r(r_1, r_2, r_3)$ of such a mixture experiment, let $f_i(r_1, r_2, r_3)$ denote the probability or probability density of (r_1, r_2, r_3) if H_i is true, $i = 1, 2$.

The conditional distributions of R_3, given $(R_1, R_2) = (r_1, r_2)$, are

$$\begin{aligned} &\text{Prob}[R_3 = r_1 \mid (r_1, r_2), H_i] = q_i; \qquad \text{and if} \quad r_2 > r_1, \\ &\text{Prob}[R_3 = r_2 \mid (r_1, r_2), H_i] = p_i = 1 - q_i, \qquad i = 1, 2, \end{aligned} \tag{4.1}$$

where $q_i = q_i(r_1, r_2)$ are determined as above by $r_1 = \log(q_2/q_1)$, $r_2 = \log(p_2/p_1)$. If $r_1 = r_2 = 0$, then $R_3 \equiv 0$, and we may take $p_1 = p_2 = 1$. Hence the marginal probability or probability density of (r_1, r_2) is

$$f_i(r_1, r_2) = \begin{cases} f_i(0, 0, 0,), & \text{if} \quad r_1 = r_2 = 0, \\ f_i(r_1, r_2, r_1)q_i + f_i(r_1, r_2, r_2)p_i, & \text{if} \quad r_1 < r_2, \end{cases} \tag{4.2}$$

for $i = 1, 2$. However $f_1(r_1, r_2) \equiv f_2(r_1, r_2)$ (a.e., H_1 and H_2), since the distribution of G of (R_1, R_2) is independent of the hypotheses. Hence we can write

$$f_i(r_1, r_2, r_3) = \begin{cases} f_1(0, 0) & \text{if } r_1 = r_2 = r_3 = 0, \\ f_1(r_1, r_2)q_i & \text{if } r_3 = r_1 < r_2, \\ f_1(r_1, r_2)p_i & \text{if } r_3 = r_2 > r_1, \end{cases} \tag{4.3}$$

for $i = 1, 2$. Hence the sufficient statistic of E_G, an arbitrary mixture $G(r_1, r_2)$ of simple binary experiments (r_1, r_2), is

$$r(x) = r(r_1, r_2, r_3) = \log [f_2(r_1, r_2, r_3)/f_1(r_1, r_2, r_3)] = r_3. \tag{4.4}$$

EXAMPLE 1.

Binomial mean. Consider the five simple binary experiments $E_0, E_1, \cdots E_4$ defined by the respective pairs of parameters (L_1, L_2) given in Table I below.

TABLE 1

Some simple binary experiments

Experiment	(p_1, p_2)	(L_1, L_2)	(α, β)
E_0	(.5, .5)	(1, 1)	(.5, .5)
E_1	(.0588, .9412)	(1/16, 16)	(.0588, .0588)
E_2	(.0039, .9961)	(1/256, 256)	(.0039, .0039)
E_3	(.0037, .9377)	(1/16, 256)	(.0037, .0623)
E_4	(.0623, .9963)	(1/256, 16)	(.0623, .0037)

Some distributions defining mixtures of the above experiments

G	G^1	$G^c, 0 \leqq c \leqq 1$
$g_0 = \binom{4}{2}(.2)^2(.8)^2 \doteq .1536$	$g_0^1 = g_0 \doteq .1536$	$g_0^c = g_0$
$g_1 = \binom{4}{1}(.2)(.8)^3 + \binom{4}{3}(.2)^3(.8) \doteq .4352$	$g_1^1 = 0$	$g_1^c = (1 - c)g_1$
$g_2 = \binom{4}{0}(.8)^4 + \binom{4}{4}(.2)^4 \doteq .4112$	$g_2^1 = 0$	$g_2^c = (1 - c)g_2$
$g_3 = 0$	$g_3^1 = (1 - g_0^1)/2 \doteq .4232$	$g_3^c = cg_3^1$
$g_4 = 0$	$g_4^1 = g_3^1 \doteq .4232$	$g_4^c = cg_4^1$

The table gives also the parameters (p_1, p_2) and (α, β) of these experiments to four decimal accuracy. The table also gives a number of discrete distributions $G^c = G^c(r_1, r_2)$: for each c, $0 \leqq c \leqq 1$, a mixture experiment E_{G^c} is defined by the five probabilities $g_i^c = \text{Prob}\ (E_i)$, $i = 0, 1, \cdots 4$. It is convenient to use the notation $g_0E_0 \oplus g_1E_1 \oplus \cdots \oplus g_4E_4$ to denote the operation of mixing the experiments $E_0, \cdots E_4$ with respective probabilities $g_0, \cdots g_4$. We can then write, for each c, $0 \leqq c \leqq 1$,

$$E_{G^c} = \sum_{i=0}^{4} \oplus\, g_i^c E_i. \tag{4.5}$$

Consider next the binomial experiment E_B consisting of four observations, with parameter $\theta = .2$ or $.8$:

$$(4.6)\quad f_1(x) = \binom{4}{x}(.2)^x(.8)^{4-x}, \qquad f_2(x) = \binom{4}{x}(.8)^x(.2)^{4-x}, \quad x = 0, 1, \cdots 4.$$

The following assertion can be verified by simple direct calculations: The mixture experiments E_{G^c} defined above are equivalent to one another, and each is equivalent to E_B. That is, $E_B = E_{G^c}$ for each c, $0 \leqq c \leqq 1$. The $v(u)$ curve of E_B is easily determined from the given binomial distributions $f_i(x)$, and consists of the line segments between the successive points (given to four-decimal accuracy): (0, 0), (.4096, .0016), (.8192, .0272), (.9728, .1808), (.9984, .5904), and (1, 1). It may be noted that only one of the above distributions G^c represents a mixture of strictly ordered simple binary experiments, namely $G^0 \equiv G$.

EXAMPLE 2.

Normal mean. The symmetric simple binary experiments (r_1, r_2) are those for which $r_1 = -r_2$. Any mixture $G(r_1, r_2)$ over this strictly ordered class of experiments can be represented conveniently by the marginal c.d.f. of R_2 under G, which we denote by $G(r_2)$. Let

$$(4.7)\qquad G(r_2) = \Phi(r_2 - \tfrac{1}{2}) - \Phi(-r_2 - \tfrac{1}{2}), \qquad \text{for} \quad 0 \leqq r_2 \leqq \infty,$$

where $\Phi(u) = \int_{-\infty}^{u} \phi(u)\, du$ and $\phi(u) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}u^2)$. Then

$$(4.8)\qquad G(r_2) = \int_0^{r_2} g(y)\, dy,$$

where

$$g(y) = \phi(y - \tfrac{1}{2}) + \phi(-y - \tfrac{1}{2}).$$

Under hypothesis H_i, the sufficient statistic r_3 of the mixture experiment E_G has the density function

$$(4.9)\qquad f_i(r_3) = \begin{cases} g(r_2)q_i(-r_2, r_2) & \text{if} \quad r_3 < 0, \\ g(r_2)p_i(-r_2, r_2) & \text{if} \quad r_3 > 0, \\ g(r_2) & \text{if} \quad r_3 = 0, \end{cases}$$

where $r_2 = |r_3|$, $q_1(-r_2, r_2) = (e^{r_2} - 1)/(e^{r_2} - e^{-r_2})$, $q_2(-r_2, r_2) = e^{-r_2}q_1(-r_2, r_2)$, and $p_i(-r_2, r_2) = 1 - q_i(-r_2, r_2)$, for $i = 1, 2$. Upon simplification we find that $f_1(r_3) = \phi(r_3 + \frac{1}{2})$, $f_2(r_3) = \phi(r_3 - \frac{1}{2})$; thus the sufficient statistic r_3 has under each hypothesis a normal distribution with unit variance, with respective means $-\frac{1}{2}$ and $\frac{1}{2}$.

Consider next the experiment E_N consisting of a single observation on a normally distributed random variable X, having unit variance and, under the respective hypotheses, means $-\frac{1}{2}$ and $\frac{1}{2}$. It is well known that for this experiment the sufficient statistic is $r(x) \equiv x$, which has under the respective hypotheses the same (normal) distributions found in the above mixture experiment E_G

for its sufficient statistic r_3. It follows that the two experiments are equivalent: $E_N = E_G$.

5. Decomposition theorem for binary experiments. In the preceding examples, two binary experiments typical of those treated in mathematical statistics were shown to be mathematically equivalent to certain mixtures of specified simple binary experiments. The following theorem shows that every binary experiment can be decomposed in this sense into simple components.

THEOREM. *Each binary experiment is equivalent to a mixture of strictly ordered simple binary experiments.*

PROOF:

1. Let $v(u)$ be an arbitrary convex c.d.f. on the closed unit interval, $v(0) = 0$, $v(1) = 1$, representing as above any given binary experiment E. E has the sufficient statistic u with distributions

$$\text{Prob}\{U \leqq u \mid H_1\} = u \equiv \int_0^u du, \qquad \text{Prob}\{U \leqq u \mid H_2\} = v(u), \qquad 0 \leqq u \leqq 1; \tag{5.1}$$

and for $u < 1$, $v(u) = \int_0^u f_2(u)du$, where $f_2(u) \equiv v'(u)$ is the right-derivative of $v(u)$.

Let $h(u) = u - v(u)$ and $h^* = \sup\{h(u) \mid 0 \leqq u \leqq 1\}$. We have $h^* > 0$, except in the case $v(u) \equiv u$, $0 \leqq u \leqq 1$, which is the uninformative experiment $(r_1, r_2) = (0, 0)$ for which the conclusion of the theorem holds trivially. Assuming $h^* > 0$, the function $h(u)$ is concave, $h(0) = h(1) = 0$, $h(u) > 0$ for $0 < u < 1$; $h(u)$ is continuous, except possibly at $u = 1$ corresponding to a possible discontinuity of $v(u)$ at $u = 1$. If $v(u)$ is discontinuous at $u = 1$, we define $h(1)$ as multiple-valued, having all values in the closed interval $[1 - v(1-), 1]$; then in all cases $h(u)$ is a graphically-continuous concave curve on the closed unit interval. The right-derivative of $h(u)$ is $h'(u) \equiv 1 - v'(u)$, for $u < 1$.

For each h, $0 \leqq h < h^*$, the equation $h(u) = h$ has two distinct roots which we designate $u_1(h) < u_2(h)$. The equation $h(u) = h^*$ is satisfied on a closed interval or at a single point u, which we designate by $u_1(h^*) \leqq u \leqq u_2(h^*)$, $u_1(h^*) \leqq u_2(h^*)$. $u_1(h)$ is continuous, convex, and strictly increasing in h, $0 \leqq h \leqq h^*$. $u_2(h)$ is continuous, concave, and nonincreasing; it is strictly decreasing in h, for $1 - v(1-) \leqq h \leqq h^*$ (that is, for $0 \leqq h \leqq h^*$, unless $v(u)$ is discontinuous), and $u_2(h) = 1$ for $1 - v(1-) \leqq h \leqq h^*$. Let $u_i'(h)$ denote the respective right-derivatives of $u_i(h)$, for $0 \leqq h < h^*$; then

$$u_i'(h) = [1 - f_2(u_i(h))]^{-1} \quad \text{for} \quad 0 \leqq h < h^*, \qquad i = 1, 2. \tag{5.2}$$

Corresponding to each h, $0 \leqq h < h^*$, we define the simple binary experiment

$$E_h\colon (r_1[h], r_2[h]) = (\log f_2(u_1(h)), \log f_2(u_2(h))). \tag{5.3}$$

Corresponding to $h = h^*$, we take $(r_1[h^*], r_2[h^*]) = (0, 0)$. These experiments are clearly strictly ordered.

Let

$$G(h) = \begin{cases} 1 - (u_2(h) - u_1(h)), & \text{for } 0 \leqq h < h^*, \\ 1, & \text{for } h = h^*. \end{cases} \tag{5.4}$$

Let $g(h) = u_1'(h) - u_2'(h)$, for $0 \leqq h < h^*$. Then $G(h) = \int_0^h g(h)dh$ for $0 \leqq h < h^*$.

2. We define the experiment E_G as the mixture $G = G(h)$ of the strictly ordered simple binary experiments $E_h : (r_1[h], r_2[h])$, $0 \leqq h \leqq h^*$. We proceed to prove that $E = E_G$, by proving that $v(u) = v_G(u)$, $0 \leqq u \leqq 1$, where $v_G(u)$ is the canonical form of E_G.

For each $h < h^*$, the simple experiment $E_h : (r_1[h], r_2[h])$ is equivalent to an experiment consisting of one observation on the random variable U_h having the following distributions:

$$\begin{aligned} \text{Prob}\,\{U_h = u_1(h) \mid H_i\} &= q_i(h) \\ \text{Prob}\,\{U_h = u_2(h) \mid H_i\} &= p_i(h) = 1 - q_i(h), \end{aligned} \tag{5.5}$$

where $q_i(h)$, $i = 1, 2$, are determined by

$$r_1[h] = \log\,[q_2(h)/q_1(h)], \qquad r_2[h] = \log\,[p_2(h)/p_1(h)].$$

For $h = h^*$, the experiment $(r_1[h^*], r_2[h^*]) = (0, 0)$ is equivalent to the trivial experiment consisting of one observation on the random variable U_{h^*} which has, under H_1 and H_2, the same uniform distribution on the interval $[u_1(h^*), u_2(h^*)]$. Let E_K be the experiment in which one observation h is taken on an auxiliary randomization variable H with the c.d.f. $G(h)$ defined above, independent of the hypotheses, followed by one observation on the corresponding random variable U_h whose distributions under H_1, H_2, were given above. Each possible outcome of this mixture experiment has the form (h, u_h) where h is the observed value of H and u_h is the observed value of U_h. Clearly $E_K = E_G$.

For different values of h, the ranges of U_h are disjoint; hence the observed value h is a function of the observed value u_h, and the latter is a sufficient statistic for E_G. The distributions of the statistic u_h are those of the random variable U_H, which are determined as follows: Let $W_i(u) = \text{Prob}\,\{U_H \leqq u \mid H_i\}$, $0 \leqq u \leqq 1$, $i = 1, 2$. We have $W_i(1) = 1$; and since Prob $\{H = 0\} = G(0) = 0$, $W_i(0) = 0$, for $i = 1, 2$. For $0 < u < u_1(h^*)$, we have

$$W_i(u) = \int_0^u w_i(u)\,du, \qquad i = 1, 2, \tag{5.6}$$

where

$$\begin{aligned} w_i(u) &= g(h(u))q_i(h(u))/u_1'(h(u)) \\ &= [u_1'(h(u)) - u_2'(h(u))]q_i(h(u))/u_1'(h(u)). \end{aligned}$$

Hence

$$w_1(u) = [1 - u_2'(h(u))/u_1'(h(u))] \cdot [f_2(u_2(h(u)) - 1]/[f_2(u_2(h(u))) - f_2(u_1(h(u)))]. \tag{5.7}$$

We have $u_1(h(u)) \equiv u$, and for brevity we write here u_2 for $u_2(h(u))$, for $0 < u < u_1(h^*)$. Thus

$$w_1(u) = (1 - [1 - f_2(u)]/[1 - f_2(u_2)]) \cdot [f_2(u_2) - 1]/[f_2(u_2) - f_2(u)] = 1. \tag{5.8}$$

Since $q_2(h(u)) = f_2(u)q_1(h(u))$, we have, for $0 < u < u_1(h^*)$,

$$w_2(u) = f_2(u). \tag{5.9}$$

In the same way the same formulae for $w_i(u)$ can be verified for the range $u_2(h^*) < u < 1$. If Prob $\{H = h^*\} = 0$, $u_1(h^*) = u_2(h^*)$, and Prob $\{U_H = u_1(h^*) \mid H_i\} = 0$ for $i = 1, 2$. If Prob $\{H = h^*\} > 0$, $u_1(h^*) < u_2(h^*)$, and by definition we have, for $u_1(h^*) \leqq u \leqq u_2(h^*)$, $w_1(u) \equiv w_2(u) \equiv f_2(u) \equiv 1$. Thus $v_G(u) = \int_0^u f_2(u)du = v(u)$ for $0 \leqq u < 1$, and $v_G(1) = v(1) = 1$, completing the proof that $E_G = E$.

B. INFERENCE METHODS WITH PROBABILISTIC JUSTIFICATIONS.

6. On the mathematical treatment of statistical inference problems. It is usual in modern mathematical statistics to restrict consideration to inference problems formulated on the basis of specified statistical experiments E in which the possible probability distributions of outcomes are described and delimited. (This includes problems of experimental design, which concern the appraisal and comparison of alternative possible experiments.) Moreover, it is now usual to consider such a specified statistical experiment to be the essential and basic frame of reference in which the relevant properties of any inference techniques must be defined and interpreted; for example, the basic properties of techniques of testing statistical hypotheses, and of related estimation techniques, are various error-probabilities, each defined directly as a probability in a specified experiment E, and interpreted in terms of relative frequencies of errors in conceptually-possible repetitions of E. Inference problems and techniques as they may be discussed outside such frames of reference are usually considered vague, and lacking in objectivity and usefulness.

The preceding sections have treated the mathematical structure of statistical experiments E in the binary case, and have left aside the remaining aspects of an inference situation, which include

(a) the conclusions or decisions among which a choice must be made on the basis of an observed outcome x of experiment E;

(b) the consequences of each possible choice, on the respective assumptions that each of the simple hypotheses is true; and

(c) the evaluations of such consequences by the individual in the inference

situation; his purposes; and possibly his prior opinions or information concerning the hypotheses.

The specification of these additional aspects of an inference situation in appropriate and formal terms is often difficult or problematical, even when all of the general features of the inference situation are quite clear.

If at least aspect (a) can be specified definitely, as for example that just two conclusions or decisions are allowed, then it is possible to give an analysis of the inference problem having general usefulness in connection with various formal or informal specifications of the remaining aspects (b) and (c).

7. Tests of statistical hypotheses; two-decision problems. If it is specified that one of just two conclusions or decisions must be adopted on the basis of an outcome of E, with specified $v(u)$, we may denote by d_1 that conclusion or decision which would be more appropriate if H_1 were true, and by d_2 the alternative, which may be called "reject H_1." Then each (Lebesgue measurable) function $d = d(u)$, taking values d_1 or d_2 only, represents a possible inference rule, whose relevant properties are the error-probabilities

$$\alpha = \alpha_d = \text{Prob}\,(d(U) = d_2 \mid H_1), \quad \text{and}$$

(7.1)

$$\beta = \beta_d = \text{Prob}\,(d(U) = d_1 \mid H_2).$$

For each α, $0 \leqq \alpha \leqq 1$, let $d_\alpha(u) = d_2$ if $u \geqq 1 - \alpha$; let $d_\alpha(u) = d_1$ if $u < 1 - \alpha$. Then the error-probabilities of $d_\alpha(u)$ are α and

$$\beta = \beta(\alpha) = \begin{cases} v([1 - \alpha]), & \text{for } 0 < \alpha \leqq 1, \\ v(1-), & \text{for } \alpha = 0. \end{cases} \tag{7.2}$$

Since the likelihood ratio statistic of E is $v'(u)$, a non-decreasing function of u, we have by the fundamental lemma of Neyman and Pearson that $d_\alpha(u)$ is a best test of H_1 against H_2 of significance level α.

Let $\alpha' = \min\,[\alpha \mid \beta(\alpha) = 0] \equiv 1 - \max\,[u \mid v(u) = 0]$. The inference functions $d_\alpha(u)$, $0 \leqq \alpha \leqq \alpha'$, constitute a minimal essentially complete class of (admissible) inference functions. For the problem considered, on the basis of the given experiment E, no other inference functions need be given consideration; but no further analysis or simplification of the problem of choosing one of these inference functions can be given except in relation to formal or informal specifications of the aspects (b) and (c) of the inference situation referred to in the preceding Section.

8. Multi-decision problems; tests based on critical levels. To illustrate most simply that even with a binary experiment it is sometimes appropriate to allow more than two possible decisions (or conclusions), consider the case in which three decisions may be allowed. Assume that decision d_1 would be the most appropriate of the three possibilities, and that d_2 would be the least appropriate,

if H_1 were true; and that d_1 would be least appropriate, and d_2 most appropriate, if H_2 were true; the remaining decision, d_3, is then more appropriate than d_2 if H_1 is true, and more appropriate than d_1 if H_2 is true. An example would be a situation of industrial acceptance sampling in which it is assumed that each lot of items contains either a certain small proportion of defective items (H_1) or a certain higher proportion of defective items (H_2); and the possible classifications are: d_1, "apparently high quality"; or d_2, "apparently low quality"; or d_3, "indeterminate quality". Another type of example is represented by designating d_1 as the conclusion "reject H_2 (in favor of H_1)," and d_2 as the conclusion "reject H_1 (in favor of H_2)," and d_3 as the conclusion "reject neither hypothesis" or "no conclusion."

Any inference procedure here can be represented by some function $d(u)$, defined on the unit interval, taking values d_1, d_2 or d_3. The relevant properties of any such function are just the four error-probabilities α_i, β_i, $i = 1, 2$, where

$$\alpha_1 = \text{Prob}\,[d(U) = d_2 \mid H_1] = \text{the probability of a "major Type I error,"}$$

$$\alpha_2 = \text{Prob}\,[d(U) = d_3 \mid H_1] = \text{the probability of a "minor Type I error,"}$$

$$\beta_1 = \text{Prob}\,[d(U) = d_1 \mid H_2] = \text{the probability of a "major Type II error," and} \tag{8.1}$$

$$\beta_2 = \text{Prob}\,[d(U) = d_3 \mid H_2] = \text{the probability of a "minor Type II error."}$$

Clearly the general goal, in appraising and selecting an inference function based on a given binary experiment, is that each of these error-probabilities should be suitably small. If the function $\beta(\alpha)$ is defined as above, then for any values of α_1 and α_2 such that $0 \leqq \alpha_1 + \alpha_2 \leqq \alpha'$ (no other cases should be considered), we have (by the Neyman-Pearson lemma) that the smallest possible value of β_1 is $\beta(\alpha_1 + \alpha_2)$, and the smallest possible value of β_2 is $\beta(\alpha_1) - \beta(\alpha_1 + \alpha_2)$; and that these are the error-probabilities of the admissible three-decision function:

$$d(u) = \begin{cases} d_1, & \text{if } u < 1 - \alpha_1 - \alpha_2, \\ d_3, & \text{if } 1 - \alpha_1 - \alpha_2 \leqq u < 1 - \alpha_1, \\ d_2, & \text{if } 1 - \alpha_1 \leqq u. \end{cases} \tag{8.2}$$

Comments like those of the preceding Section apply to the problem of choice of a particular inference function of this form. Any inference or decision function of this form has the *probabilistic justification* that its four error-probabilities are "jointly minimum" in the sense that no one of them could be reduced except by an increase in one or more of the others. The *policy* of using such an inference or decision function, having suitably small error-probabilities, is thereby justified in the sense that in many independent applications, under respective hypotheses, the relative frequencies of the more and less serious errors of various kinds will tend to be correspondingly small.

The preceding discussion can be immediately generalized to allow any number of possible decisions or conclusions, simply ordered according to their decreasing appropriateness if H_1 is true (and increasing appropriateness if H_2 is true); an infinite number (not necessarily countable) can be allowed. In all such cases, the admissible inference or decision functions, having probalistic justifications of the kind illustrated above, will have a form in which larger values of the outcome u tend to indicate conclusions or decisions which are more appropriate when H_2 is true.

An inference technique which antedates modern mathematical statistics, and which remains in wide use, is that based on the *critical level* associated with an observed outcome: When an appropriate statistic has been selected, for example the statistic u, the critical level is defined as the probability, under a hypothesis H_1 being tested, of a value of U at least as large as the value observed:

$$\alpha(u) = \text{Prob}\,[U \geqq u \mid H_1]. \tag{8.3}$$

Observed values of $\alpha(u)$ more or less close to 0 are customarily interpreted as representing more or less strong evidence for rejection of H_1; one convention of interpretation, which is clearly rather schematic, applies the term "significant" to outcomes $\alpha(u) \leqq .05$, and the term "highly significant" to outcomes $\alpha(u) \leqq .01$. Leaving aside interpretations which ascribe to a numerical value of $\alpha(u)$ some intrinsic meaning as a quantitative measure of strength of evidence against H_1 in an outcome u, there remains the qualitative simple ordering of conclusions with those favoring H_2 more strongly corresponding to smaller values of $\alpha(u)$.

This latter qualitative part of the customary interpretation of various possible values of the critical level, considered in the context of a specified experiment, has the kind of probabilistic and frequency justification described above. In addition, the numerical values of $\alpha(u)$ have probabilistic interpretations related to various errors of Type I; for example, any interpretation of outcomes $\alpha(u) \leqq .01$ as "strong evidence against H_1" will be highly inappropriate if H_1 is true, but will be made with probability only .01 when H_1 is true. However techniques based upon critical values do not incorporate systematic consideration of error-probabilities under H_2.

While the theory of Neyman and Pearson introduced the essential complementary concept of errors of Type II, the formal development and the applications of this theory have typically been based on fixed-level formulations, and have typically treated only two-decision problems. The preceding discussion shows that a simple adaptation of the standard fixed-level theory and methods gives multi-decision and corresponding inference methods which have the flexibility and intuitive appeal of the traditional critical level technique, and also an appropriately complete objective probabilistic appraisal and justification based on consideration of probabilities of errors of all kinds and degrees, in the context of a specified statistical experiment.

C. INFERENCE METHODS WITH INTRINSIC JUSTIFICATIONS.

9. Informative inference. A traditional and basic type of application of techniques of mathematical statistics, including techniques described in the preceding two sections, occurs in situations of empirical scientific research. In such situations, besides problems of inference or decision-making which bear upon specific practical purposes, or specific research purposes such as drawing working conclusions and planning further research, a broader inference problem is often recognized and dealt with. The latter problem is that of recognizing, appraising, and sometimes reporting in the scientific or technical literature, in appropriate objective terms, the general character of experimental results as they are relevant to statistical hypotheses (or values of unknown parameters) of interest. This problem may be described as that of recognizing, and reporting appropriately, *statistical evidence* relevant to statistical hypotheses of interest. For brevity, we use the term *informative inference* to refer to this problem and to methods for dealing with it.

In typical research situations, when a test of a statistical hypothesis (appropriately valid and efficient) indicates rejection of that hypothesis, besides the conclusions or decisions which the experimenter may reach it is often recognized that the experimental results may be of more general interest and value; and a description of the testing procedure and its outcome are often reported to indicate in objective terms the character of the results as evidence relevant to hypotheses. The reporting of estimates of parameters of interest with indicators of their precisions, in the scientific literature, typically serves the same broad and basic scientific function. In this function, the methods of mathematical statistics serve as techniques for the *evidential interpretation* of experimental outcomes.

The basic terms of such interpretations are usually taken to be certain error-probabilities associated with the testing or estimation techniques used. (The precision of an estimator can typically be interpreted in terms of probabilities of estimation-errors of various magnitudes.) In fact the general nature of statistical evidence, relevant to hypotheses of interest, is commonly recognized, expressed, and dealt with, in a generally clear and effective way, in terms of such error-probabilities. Our purpose in the following sections is to clarify the mathematical structure of statistical evidence and the terms appropriate for its description.

10. Symmetric simple binary experiments. It is convenient to refer to the outcome r_2 of any simple binary experiment (r_1, r_2) as "positive," and to the outcome r_1 as "negative." A simple binary experiment will be called symmetric if $r_1 = -r_2$, that is, if the experiment is of the form $(-r_2, r_2)$; in the present section we consider only experiments of this form. Each such experiment is characterized by a number r_2, $0 \leqq r_2 \leqq \infty$. This class of experiments is *simply* ordered, by the parameter r_2, according to the relation "more informative than" defined in Section 3 above.

There is no difficulty in recognizing the appropriate evidential interpretations of outcomes of the extreme cases in this class of experiments. The completely informative experiment $(-\infty, \infty)$ gives outcomes each of which can naturally be called completely informative: the outcome $r = \infty$ supports the certain inference that H_1 is false and H_2 is true. An alternative interpretation, which is equivalent for all purposes of application, is: the inference that H_2 is true is practically certain, in the highest possible degree. Similarly, the outcome $r = -\infty$ supports the certain inference that H_1 is true. The uninformative experiment (0, 0) gives outcomes each of which can naturally be called (completely) uninformative: an outcome $r = 0$ has no relevance to the hypotheses, and therefore gives no support in any degree to any inferences concerning the hypotheses.

In any intermediate case $(-r_2, r_2)$, $0 < r_2 < \infty$, it is natural and necessary to attribute to the positive outcome the qualitative evidential property of *supporting* H_2 (as against H_1), and to the negative outcome the property of supporting H_1. In addition to intrinsic plausibility, these qualitative evidential properties attributed to the possible numerical values of r, r_2 or $-r_2$, have the objective interpretation and justification that, under each hypothesis, the probability that such an interpretation will be qualitatively inappropriate (the probabilities of a "false positive" (Type I error) and of a "false negative" (Type II error), in the obvious simplest testing or two-decision procedure) is equal to

$$\alpha = \alpha[r_2] = 1/(1 + e^{r_2}) < \tfrac{1}{2}. \tag{10.1}$$

If $0 < r_2 < r_2' < \infty$, we interpret the positive outcome of the experiment $(-r_2', r_2')$ as *supporting H_2 more strongly than* the positive outcome of the experiment $(-r_2, r_2)$. This interpretation is supported by the considerations that outcomes statistically equivalent to those of the latter experiment can be generated by modifying outcomes from the former experiment by the "addition of pure noise" unrelated to the hypotheses, in the sense of Section 3 above; and that $\alpha[r_2'] < \alpha[r_2]$, since $\alpha[r_2]$ decreases from $\frac{1}{2}$ to 0 as r_2 increases from 0 to ∞.

In summary, over the class of symmetric simple binary experiments, the function $r = \log [f_2(x)/f_1(x)]$ has been given an unequivocal and consistent set of evidential interpretations: $r = r(x)$ *is an objective, internally-consistent and efficient indicator of evidence relevant to hypotheses in experimental outcomes.*

11. Symmetric binary experiments. A binary experiment E, not necessarily simple, will be called *symmetric* if its canonical form $v(u)$ is symmetric about the line $u + v = 1$; that is, if for each u, $0 \leqq u \leqq 1$, we have $v(1 - v(u)) = 1 - u$. For any such experiment, the method of the proof of the decomposition theorem of Section 5 above gives a mixture experiment, equivalent to E, each of whose simple components has the symmetric form $(-r_2, r_2)$; as in Example 2 of Section 4 above, any such mixture can be represented by a (generalized) c.d.f. $G(r_2)$ on the range $0 \leqq r_2 \leqq \infty$. For any given symmetric binary experiment E, let E_G denote this equivalent mixture experiment.

Since E_G and E are mathematically equivalent, in particular for purposes of informative inference, and related questions of evidential interpretations of out-

comes, we can consider any outcome r of E as if it were a mathematically-corresponding outcome of E_G. Each outcome of this symmetric mixture experiment E_G has the form $(-r_2, r_2, r)$, where $r = r_2$ or $-r_2$. Since r_2 is the observed value of a random variable having under each hypothesis the same known distribution $G(r_2)$, the observed value r_2 is irrelevant as evidence concerning the hypotheses. The observed value r_2 determines the symmetric simple binary experiment $(-r_2, r_2)$ which is performed; hence $r_2 = |r|$ indicates, as in the preceding Section, just the strength of the evidence which is provided by the outcome r of the experiment $(-r_2, r_2)$. It is possible and necessary to interpret the outcome r of the latter experiment in the way established in the preceding section for outcomes of symmetric simple binary experiments, for purposes of informative inference, since the appropriate frame of reference for considering the evidential character of r is clearly the selected simple experiment, and the structure of E_G is otherwise clearly irrelevant to such interpretations.

Because of the equivalence of E and E_G, and the related equivalence between outcomes of the two respective experiments having numerically equal values r, we obtain from the preceding paragraph the following general conclusions: *Given any symmetric binary experiment E, for purposes of informative inference, any outcome r of E must be interpreted evidentially in the same way as a numerically-equal outcome of a symmetric simple binary experiment. In particular, given r, the mathematical form of E is irrelevant for such purposes and interpretations.*

To illustrate this conclusion in concrete terms, a physical interpretation of Example 1 of Section 4 above may be useful. Suppose that four measurement instruments (or techniques of observation) are available in an investigation concerning two hypotheses, with each instrument giving dichotomous outcomes "positive" or "negative," and each instrument symmetric in the sense that it has equal probabilities α of false positives and of false negatives. Let the simple experiments E_0, E_1, E_2 defined in Example 1 represent respectively three of these instruments, when each is used without replication (to obtain a single observation). Let the fourth instrument have $\alpha = .2$, and let E denote the experiment consisting of four independent measurements by this instrument; then E is the binomial experiment of Example 1.

Let E_G denote an experimental procedure in which one of the first three instruments is selected at random, with the respective probabilities g_i given in the Example, and in which the instrument selected is used to obtain a single measurement. With this procedure, if the worthless instrument E_0 happens to be selected, one may fairly plead victimization by rather improbable bad luck, and indeed one had good reason to hope for and expect selection of a more informative instrument; however these considerations are irrelevant to the problem of making informative inferences from a measurement provided by E_0 to the hypotheses; for this problem, the only relevant considerations are that the instrument and its measurements are strictly worthless, and that this outcome of the experiment E_G provides, recognizably, no contribution whatsoever to the inference problem.

In terms of the binomial experiment E, the outcome $x = 2$ corresponds (under the mathematical equivalence of E with E_G) to the selection of E_0 (and occurrence of either of its outcomes) in E_G. Hence there is no reason to give the binomial outcome, $x = 2$, interpretations differing in any respect from the interpretations just described for the outcome E_0 of E_G. Nor is there any reason to consider any other aspect of the binomial model of the experiment E, for purposes of informative inference, given that $r(x) = r(2) = 0$, a recognizably (completely) uninformative outcome.

Suppose, alternatively, that in the mixture experiment E_G the most informative instrument, E_2, is by good fortune selected. Granting that the occurrence of such good luck is irrelevant as evidence regarding the hypotheses, it is most relevant to the quality or strength of inferences which may be made from a measurement supplied by E_2. Evidently there is no reason to qualify or weaken the resulting inference statements on the ground that one was not sure beforehand that one would have the good luck to be able to use the best possible instrument. Suppose that use of the selected instrument E_2 gives a positive outcome, $r = 256$. Under the mathematical equivalence between E and E_G, this outcome corresponds to the outcome $x = 4$ of the binomial experiment E (that is, to four positive outcomes in four independent measurements by the instrument having $\alpha = .2$), for which we also (necessarily) have $r(x) = 256$. It follows that the outcome $x = 4$ of the binomial experiment E should be interpreted in exactly the same way, as evidence relevant to the hypotheses, as if it were a positive outcome obtained in a single measurement by an instrument E_2 having probability $\alpha \doteq .0039$ of false positives and of false negatives. The numerical value $r = \log (1 - \alpha)/\alpha = \log 256$ serves, by definition, as a compact abbreviation for such an evidential interpretation of the outcome $x = 4$. Analogous interpretations apply to the remaining possible outcomes x of E.

12. Binary experiments in general. To extend the scope of the preceding evidential interpretations of the statistic r to binary experiments which are not necessarily symmetric, let E: $v(u)$ be any binary experiment. Let E^*: $v^*(u)$ be the "reflection" of $v(u)$ in the line $u + v = 1$; that is, for each point $(u', v(u'))$ of the (continuous) graph of $v(u)$, let the graph of $v^*(u)$ contain the point $(u'', v^*(u'')) \equiv (1 - v(u'), 1 - u')$. Let $E^{**} = \frac{1}{2}E \oplus \frac{1}{2}E^*$; that is, E^{**} is the mixture experiment having E and E^* as components with probabilities each $\frac{1}{2}$. Then E^{**} is a symmetric binary experiment. If the experiment E^{**} were under consideration, and if its component E were selected, then any outcome r of E must be interpreted evidentially in the way described in the preceding Section, since E^{**} is symmetric; the selection of E is irrelevant here, given the numerical value of r.

Returning to consideration of the given experiment E, any outcome r of E is equivalent, for purposes of informative inference, to an outcome of the mixture experiment E^{**} in which the component E is first selected, and then the outcome r is observed. It follows that the evidential interpretations of outcomes r of *any* binary experiment must be of the same kinds as those given in the cases discussed in the preceding Section.

13. Inferences based on the likelihood function. The results of the preceding analysis may be summarized as follows: When any binary experiment E is used for purposes of informative inference, and when any specified outcome r of E is obtained, the mathematical structure of E is then irrelevant to those purposes, and just the numerical value r is relevant. Any such observed numerical value r has an intrinsic objective probabilistic character as evidence relevant to H_1 or H_2 ; namely: (a) the qualitative property that the outcome favors H_2 if r is positive, favors H_1 if r is negative, and is irrelevant as evidence if $r = 0$; and (b) strength, as evidence, identical with that of a single outcome of the symmetric simple binary experiment $(-r_2, r_2)$, where $r_2 = |r|$. The latter simple experiment has probabilities of false positives and of false negatives each equal to

$$\alpha = \alpha[|r|] = 1/(1 + e^{|r|}) \leqq \tfrac{1}{2}. \tag{13.1}$$

We may say that such inferences are based just on the likelihood function $[f_1(x), f_2(x)]$ on the observed outcome x, since $r(x)$ is a compact representation of the likelihood function in the case of any binary experiment.

If any evidential interpretations of observed values of $r(x)$ are regarded within the frame of reference of the specific binary experiment E from which x is obtained, then we have formally a particular case of the procedures discussed in Section 8 above. However, such evidential interpretations of outcomes $r(x)$, despite their objective aspects, are in general deficient for purposes of informative inference to the extent that they differ from the evidential interpretations of the likelihood function described above.

14. Appraisal and design of experiments for informative inference. Granting that the structure of a binary experiment is irrelevant to the evidential interpretation of an outcome x, apart from determination of $r(x)$, there remain the important problems of appraising, comparing, and designing experiments for purposes of informative inference. Here the structure of an experiment is most relevant, and the partial ordering discussed above is basic: Error-probability curves $\beta(\alpha)$ (and their analogues in more complicated experiments) which have been studied extensively in modern mathematical statistics, although usually given other interpretations, are of direct use for such purposes. No simple ordering of experiments, nor numerical measure of information in outcomes or experiments, seems adequate for such purposes in general (although possibly useful in a large-sample approximate sense), since the evidential meanings and values of numerical values $r(x)$ are primitive (although objective) and the distributions of $r(X)$ can in principle be considered directly.

As an example of experimental design problems for informative inference, suppose that for two simple hypotheses it is required to obtain as economically as possible statistical evidence with strength represented by $|r(x)| \geqq \log 99$. If repeated independent observations Y_i are available, with densities $g_1(y)$, $g_2(y)$ under the respective hypotheses, and if costs depend only upon the number of observations (increasing with the latter in any way), it follows immediately that the most economical experimental design is given by the sequential sampling

rule which terminates when for the first time the observations taken, $x = (y_1, \cdots y_n)$, satisfy $|r(x)| \geqq \log 99$. Such sampling rules are the same as Wald's, given for the problem of sequential testing between two simple hypotheses. (The elementary determination of this rule as best for informative inference contrasts sharply with the difficult proof of its optimality for the testing problem.) If indefinitely large sample sizes are not allowed, even with small probability, the specification of the problem must be altered.

15. Relations between statistical evidence and significance tests. Let E be any informative binary experiment, $v(u) \not\equiv u$, and for some α, $0 < \alpha < \alpha'$, let $d_\alpha(u)$ be the best test of level α as defined in Section 8 above. Then as above this test has $\beta = 1 - v(1 - \alpha)$, $0 < \beta < 1$. If outcomes of E are reported only in the form, either d_2 : "reject H_1"; or d_1 : "do not reject H_1" (or "accept H_1"), then this significance test procedure is equivalent to the simple binary experiment E' in which the likelihood ratio statistic L has only the two possible values $L_1 = \beta/(1 - \alpha) < 1$ (for d_1) or $L_2 = (1 - \beta)/\alpha > 1$ (for d_2). Hence the outcome "reject H_1" has strength, as evidence, corresponding to the value L_2 of the likelihood ratio statistic, and is associated intrinsically in the sense of Section 14 above with the error-probability $\alpha^* = 1/(1 + L_2) = \alpha/(1 - \beta + \alpha)$.

If the ratio L_2 of the test's power $(1 - \beta)$ to its level α is not far above unity, then α^* is not far below .5, and the evidential strength of the outcome "reject H_1" is correspondingly slight; this can be the case within wide limits, for any value of α, including very small values. Thus in binary experiments *a small value of α does not in general imply high evidential strength in the outcome "reject H_1"*, and the determination of the evidential strength of such an outcome depends upon β as well as α, through the function $L_2 = (1 - \beta)/\alpha$. (Within a specified binary experiment, if α is decreased, then L_2 is increased, at least if $v(u)$ is strictly convex; however the upper limit approached by L_2 as α decreases may or may not be far above unity.)

On the other hand, if β is appreciably below .5, then small values of α correspond to similarly small values of α^*, the error-probability intrinsically associated with the outcome "reject H_1 ." For example, $\beta < .25$ implies $\alpha/(1 + \alpha) < \alpha^* < (4/3)\alpha$; if α is also small, such inequalities imply that $\alpha^* \doteq \alpha$. That is, *if both α and β are small, then the error-probability α^* corresponding to the intrinsic evidential strength of the outcome "reject H_1" is approximately equal to α.*

Parallel remarks apply to evidential interpretations of the outcome "accept H_1 ."

While the preceding considerations clarify, and in important cases support, certain qualitative and quantitative features of customary uses and interpretations of significance tests as techniques for informative inference, they do not completely support the method of significance tests as such for purposes of informative inference. For such purposes, the methods based as described above directly on the likelihood function are preferable in principle, for the reasons given there.

16. Relations of statistical evidence to prior information and to conclusions. The preceding sections have dealt with a single aspect of situations of informative inference: the nature and properties of experimental outcomes as evidence relevant to statistical hypotheses. If each statistical hypothesis represented in a binary experiment is regarded initially as possibly true, then in many situations evidence against one hypothesis, if sufficiently strong, would support a *conclusion* that that hypothesis is false. The general nature of conclusions in various contexts of investigation, their uses, limitations, and possible ultimate reversibility, are familiar (cf., Tukey, [4]). These features of conclusions, and the strength of statistical evidence which would suffice in any given situation to support a conclusion, are among the aspects of inference situations (like (b) and (c) of Section 9 above) whose formal specification is problematical. But the process by which informal consideration of the various aspects of inference situations, including experimental outcomes, sometimes leads to conclusions, is familiar; and the formal and objective evidential properties of experimental outcomes, analyzed above, are conveniently assimilable in this process.

One aspect of an inference situation whose formal specification is often problematical is that of prior opinions or information, including relevant previous experience, indirect evidence, and general theoretical considerations. Bayesian treatments of inference problems, in which such considerations are represented by prior probabilities (in some sense) of the statistical hypotheses considered, will not be discussed here, except to note that they coincide with the informal process referred to in the preceding paragraph in taking just the likelihood function as the appropriate indicator of evidence in outcomes relevant to the hypotheses, and that they differ only in their degree and mode of formalization of other aspects of an inference situation.

17. Acknowledgments. An example given by Cox [5] illustrated the usefulness of mixtures of experiments for analysis of problems in the foundations of statistical inference. A special status and role of the likelihood function in informative inference was pointed out by Fisher and by Barnard [6]; however the above methods of analysis and interpretation are new.

REFERENCES

[1] BOHNENBLUST, H. F., SHAPLEY, L. S. AND SHERMAN, S., "Reconnaissance in game theory," Research Memorandum RM-208, The Rand Corporation, Santa Monica, August 12, 1949.

[2] BLACKWELL, DAVID, "Comparison of experiments," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 1951, pp. 93–102.

[3] SHANNON, CLAUDE E., "A note on a partial ordering for communication channels," *Information and Control*, Vol. 1 (1958), pp. 357–372.

[4] TUKEY, JOHN W., "Conclusions vs. decisions," *Technometrics*, Vol. 2 (1960), pp. 423–433.

[5] COX, D. R., "Some problems connected with statistical inference," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 357–372.

[6] BARNARD, G. A., "Statistical inference," *J. Roy. Stat. Soc.*, Suppl., Vol. 11 (1949), pp. 115–139.