The Annals of Statistics
1973, Vol. 1, No. 6, 1027–1045

PAIRED COMPARISONS WITH ORDER-EFFECTS

BY WILLEM SCHAAFSMA

University of Groningen

Suppose n tea-tasting ladies are asked to compare two brands of brandy. Each lady drinks two cups, one of each brand. The order is determined by coin-tossing. The lady is forced to state whether she prefers the first or the second cup. Thus each lady provides one out of four possible outcomes: which brand did she try first, did she prefer the first or the last cup. One usually summarizes this information by telling for each lady which of the brands she preferred. This paper shows that by doing so some relevant information is lost. One obtains a theory for a 2×2-table where no UMP unbiased test exists. Several modifications and generalizations are considered.

1. Introduction and summary. Aiming at an "efficient design" for comparing two treatments, say a genuine treatment and a placebo, one often uses the individuals as their own control: both treatments are assigned to each individual. Sometimes there is no difficulty with respect to the order in which both treatments are assigned because they are assigned at the same time: a housewife is asked to state which of two packages of a certain product she likes best. In this paper we are interested in situations where both treatments are assigned in a certain order which may have an effect upon the outcome. There may exist a learning or an adaptation effect: the second glass of Dutch gin tastes better than the first glass, most of the time and no matter the brand. The order-effects have to be "counterbalanced" by the design.

One often uses the coin-tossing design. This is a randomized design where the order is determined by tossing a fair coin for each individual separately, "head" indicating that the genuine treatment is tried last and "tail" that the placebo is tried last. This paper will show that the usual techniques for treating the data thus obtained tacitly apply *a reduction of the data where relevant information is removed*. How to use this relevant information is another, very complicated, problem. A detailed analysis will be given (Sections 2—7) of the simple situation where each individual ranks the two treatments: the individual is forced to state whether he prefers the treatment tried first or that tried last. One usually applies the sign test. The outcomes are determined of the rv's $S_1, \cdots, S_n$ where $S_i = 1$ if the ith individual prefers the genuine treatment (the ith tossing provided "head" and the ith individual preferred the last treatment, or "tail" appeared and the first treatment was preferred), $S_i = 0$ otherwise. Next the sign test is carried out by rejecting for large values of $S = S_1 + \cdots + S_n$. This test

Received September 1970; revised September 1972.

AMS 1970 *subject classifications*. Primary 62F05; Secondary 62B99.

Key words and phrases. A 2×2 table where no UMP unbiased test exists, an incomplete necessary and sufficient statistic.

is obviously UMP among all tests based on $(S_1, \cdots, S_n)$. But $(S_1, \cdots, S_n)$ is not a sufficient statistic.

Reduction by means of a necessary and sufficient statistic (Section 3) leads to a new problem for a 2×2 table. Contrary to the classical situations, there now does not exist a UMP unbiased size-α test (if $2^{-n} < \alpha < 1 - 2^{-n}$). One can consider some particular tests like the sign test and Fisher's exact test (Sections 4 and 5) though they may be inadmissible (Section 7). One can also construct the class of all SMP unbiased size-α tests (Section 6). In practice one will use Fisher's exact test if n is large and the sign test if n is small (Section 5).

Apart from the coin-tossing design, some other designs will be considered (Sections 8, 9 and 10). Here Fisher's exact test will always turn out to be UMP unbiased size-α. Also other modifications and generalizations can be considered (Section 11).

2. The precise formulation of the testing problem. Consider the following classical experiment for comparing a drug and a placebo. Each of n individuals is asked to compare the effects of two drugs which are administered to him in an order which is determined by tossing a fair coin for each individual separately (this is the *coin-tossing design*; in Sections 8, 9 and 10 other designs will be considered). The individual is forced to state whether he prefers the first drug tried or the second one (other possibilities will be discussed in Section 11). By performing the experiment it becomes known for each of the n individuals (i) whether he tried the genuine drug first or the placebo, (ii) whether he preferred the first drug tried or the last. Accordingly the outcome space $\mathscr{X}$ of our experiment may be described as the space of all 2^{2n} possible sequences $x = (x_1, \cdots, x_{2n})$ of zeros and ones, $x_{2i-1} = 0$ (or 1 respectively) if the ith individual tried the placebo (genuine drug) first, $x_{2i} = 0$ (or 1) if the ith individual preferred the drug tried first (or last) $(i = 1, \cdots, n)$.

It is convenient to introduce the coordinate representing random variables $X_i: \mathscr{X} \to \{0, 1\}$, defined by $X_i(x) = x_i$ if $x = (x_1, \cdots, x_{2n})$. If the n individuals may be regarded as successive drawings from the "large" population under investigation, then one will accept the probabilistic assumption that the n random vectors (X_{2i-1}, X_{2i}) $(i = 1, \cdots, n)$ are independently and identically distributed. We have to characterize the joint distribution of X_{2i-1} and X_{2i}. Introduce the parameters

$$\theta_j = P(X_{2i} = 1 \mid X_{2i-1} = j) \qquad (j = 0, 1) \tag{2.1}$$

where θ_0 is the conditional probability that a randomly selected individual prefers the drug tried last (here the genuine drug) if he tried the placebo first; θ_1 is the conditional probability to prefer the drug tried last (now the placebo) if the genuine drug is tried first. Now the probabilistic model can be defined. To the value $\theta = (\theta_0, \theta_1)$ of the unknown parameter corresponds the probability distribution P_θ over $\mathscr{X}$ which is determined by the probabilities

$$P_\theta(\{x\}) = 2^{-n} \prod_{i=1}^{n} P_\theta(X_{2i} = x_{2i} \mid X_{2i-1} = x_{2i-1}) \tag{2.2}$$

of the elementary events $x = (x_1, \cdots, x_{2n})$. The product in the right-hand side consists of factors θ_0, θ_1, $1 - \theta_0$ and $1 - \theta_1$.

The meaning of θ_0 and θ_1 makes clear that we shall be interested in testing the null hypothesis H: $\theta_0 = \theta_1$ of no difference between drug and placebo, against the *one-sided* alternative A: $\theta_0 > \theta_1$ that the genuine drug is preferred over the placebo. The parameter space Θ is the set of all $\theta = (\theta_0, \theta_1)$'s satisfying $0 \leqq \theta_1 \leqq \theta_0 \leqq 1$. The partitioning $\Theta = \Theta_0 \cup \Theta_1$ is such that Θ_0 corresponds with the diagonal $\theta_0 = \theta_1$. For a deeper-going discussion, see the beginning of Section 8.

We restrict our attention to this testing problem (H, A), though other problems are also of interest: (i) the corresponding *two-sided* problem, (ii) the problem to test the null hypothesis $\theta_0 + \theta_1 = 1$ of *no order-effects*.

3. A necessary and sufficient statistic. By introducing the random variables

$$S_{jh} = \sum_{i=1}^{n} (|X_{2i-1} - j| - 1)(|X_{2i} - h| - 1) \qquad (j, h = 0, 1) \tag{3.1}$$

we can count the numbers of different factors appearing in (2.2). By using (2.1) we get

$$P_\theta(\{x\}) = 2^{-n}(1 - \theta_0)^{S_{00}(x)}\theta_0^{S_{01}(x)}(1 - \theta_1)^{S_{10}(x)}\theta_1^{S_{11}(x)} \tag{3.2}$$

and $(S_{00}S_{01}S_{10}S_{11})$ is a sufficient statistic for our family $\{P_\theta; \theta \in \Theta\}$. It follows from $\sum\sum S_{jh} = n$ that $(S_{00}S_{01}S_{10})$ is an equivalent sufficient statistic (equivalence means that the same partitioning of $\mathscr{X}$ is introduced). We apply reduction by sufficiency and restrict our attention to *randomized* tests of the form $\varphi(S_{00}S_{01}S_{10}S_{11})$ where φ is a function with all possible outcomes of $(S_{00}S_{10}S_{10}S_{11})$ as domain and with range $[0, 1]$. Remark that φ defines a test over the original outcome space $\mathscr{X}$ as follows

$$\{\varphi(S_{00}S_{01}S_{10}S_{11})\}(x) = \varphi(S_{00}(x), S_{01}(x), S_{10}(x), S_{11}(x)) .$$

Thus our attention is restricted to tests which only depend on the following 2×2 table where the outcomes of $Y = S_{00} + S_{10}$ and $Z = S_{00} + S_{01}$ are also mentioned. It follows from elementary probability theory that $(S_{00}S_{01}S_{10}S_{11})$ has the multinomial $M\{n; \frac{1}{2}(1 - \theta_0), \frac{1}{2}\theta_0, \frac{1}{2}(1 - \theta_1), \frac{1}{2}\theta_1\}$ distribution in case $\theta = (\theta_0, \theta_1)$. Of course $Z = S_{00} + S_{01}$ has the binomial $B(n, \frac{1}{2})$ distribution, hence $E_\theta(S_{00} + S_{01} - \frac{1}{2}n) = 0$ for all $\theta \in \Theta$ and $(S_{00}S_{01}S_{10})$ is not a *complete* sufficient statistic for the family $\{P_\theta; \theta \in \Theta\}$.

TABLE 1

	Placebo first	Drug first	Total
First preferred	$S_{00}(x)$	$S_{10}(x)$	$Y(x)$
Last preferred	$S_{01}(x)$	$S_{11}(x)$	$n - Y(x)$
Total	$Z(x)$	$n - Z(x)$	n

LEMMA. $(S_{00}S_{01}S_{10}S_{11})$ *is necessary and sufficient for* $\{P_\theta;\ \theta \in \Theta\}$.

PROOF. We must show that the partitioning of $\mathscr{X}$ introduced by $(S_{00}S_{01}S_{10}S_{11})$ is the coarsest of all sufficient partitionings. Standard arguments for multiparameter exponential families ([2], page 134) infer minimality from completeness ([6], page 160). We need other arguments because even $(S_{00}S_{01}S_{10})$ is not complete. Dynkin [1] shows that one of the necessary and sufficient statistics is the mapping $\lambda: \mathscr{X} \to \mathscr{F}(\Theta)$ where $\mathscr{F}(\Theta)$ is the space of all functions over Θ and $\lambda(x) = \lambda_x$ is the *likelihood-ratio function* for which $\lambda_x(\theta) = P_\theta(\{x\})/P_{\theta_0}(\{x\})$. We take $\theta_0 = (\frac{1}{2}, \frac{1}{2})$, with the result that $\lambda_x(\theta) = 2^{2n}P_\theta(\{x\})$. All that we have to do is to show that the partitioning introduced by $(S_{00}S_{01}S_{10}S_{11})$ is coarser; or in other words that $P_\theta(\{x\}) = P_\theta(\{x'\})$ for all $\theta \in \Theta$ implies that $S_{jh}(x) = S_{jh}(x')$ $(j, h = 0, 1)$. This is elementary mathematics.

4. Description of two tests. The classical solution to our testing problem is the sign test φ_{sign} which rejects for large outcomes of $S = S_{01} + S_{10}$. This test statistic S describes the number of individuals that preferred the genuine drug; S has the binomial $B\{n; \frac{1}{2}(1 + \theta_0 - \theta_1)\}$ distribution if $\theta = (\theta_0, \theta_1)$. Obviously, φ_{sign} is UMP *among all tests based on* S. Unfortunately S is not sufficient for our family $\{P_\theta;\ \theta \in \Theta\}$ and this optimum property of the sign test is not compelling. We shall need a precise definition of φ_{sign}. This could be given (i) as a function of $x \in \mathscr{X}$, (ii) as a function of the outcome of $(S_{00}S_{01}S_{10}S_{11})$, (iii) as a function of the outcome of S. We use (ii) in order to get agreement with Section 3. Let B be a rv having the binomial $B(n, \frac{1}{2})$ distribution; let $b_{n,\alpha}$ be the smallest integer b such that $P(B \geqq b + 1) < \alpha$ and define

$$\gamma_{n,\alpha} = \{\alpha - P(B \geqq b_{n,\alpha} + 1)\}/P(B = b_{n,\alpha})\,. \tag{4.1}$$

The sign test can then be defined by

$$\begin{aligned} \varphi_{\text{sign}}(s_{00}s_{01}s_{10}s_{11}) &= 0 && \text{if} \quad s_{01} + s_{10} < b_{n,\alpha} \\ &= \gamma_{n,\alpha} && \text{if} \quad s_{01} + s_{10} = b_{n,\alpha} \\ &= 1 && \text{if} \quad s_{01} + s_{10} > b_{n,\alpha}\,. \end{aligned} \tag{4.2}$$

The 2×2 table in Section 3 suggests to apply Fisher's exact test φ_{hyp} which rejects for small outcomes of S_{00}. By studying the conditional distribution of S_{00} under the condition $Y = y;\ Z = z$, both for $\theta \in \Theta_0$ and for arbitrary $\theta \in \Theta_1$, it can be shown that φ_{hyp} is *unbiased size-*α (this can also be deduced from Section 8: φ_{hyp} is UMP unbiased size-α for the modified problem which is obtained by no longer assuming that the coin is fair) and moreover, that φ_{hyp} is UMP among all conditional level-α tests based on the condition $Y = y;\ Z = z$. This optimum property is not compelling either: it is not true that φ_{hyp} is UMP unbiased size-α. We need a precise definition of φ_{hyp} as a function of the outcome $(s_{00}s_{01}s_{10}s_{11})$. Let $H_{y,z,n}$ be a rv having the hypergeometric distribution with

$$P(H_{y,z,n} = h) = \binom{z}{h}\binom{n-z}{y-h}/\binom{n}{y} \tag{4.3}$$
$$(h = \max(0, y + z - n), \cdots, \min(y, z))\,;$$

let $h_{y,z,n,\alpha}$ be the largest integer h with $P(H_{y,z,n} \leqq h - 1) < \alpha$ and define

$$\gamma_{y,z,n,\alpha} = \{\alpha - P(H_{y,z,n} < h_{y,z,n,\alpha})\}/P(H_{y,z,n} = h_{y,z,n,\alpha}) . \tag{4.4}$$

Fisher's exact test φ_{hyp} can then be defined as follows

$$\begin{aligned} \varphi_{\text{hyp}}(s_{00}\, s_{01}\, s_{10}\, s_{11}) &= 0 && \text{if} \quad s_{00} > h_{s_{00}+s_{10},s_{00}+s_{01},n,\alpha} \\ &= \gamma_{s_{00}+s_{10},s_{00}+s_{01},n,\alpha} && \text{if} \quad s_{00} = h_{s_{00}+s_{10},s_{00}+s_{01},n,\alpha} \\ &= 1 && \text{if} \quad s_{00} < h_{s_{00}+s_{10},s_{00}+s_{01},n,\alpha} . \end{aligned} \tag{4.5}$$

5. A numerical comparison of the power properties of the two tests of Section 4. Figure 1 considers the case $n = 30$; $\alpha = .0493686$ (this value of α gives $\gamma_{n,\alpha} = 1$, see (4.1)). In Figure 1, lines of constant power are drawn for φ_{sign} and for φ_{hyp}. The straight lines belong to φ_{sign} (see the beginning of Section 4). Figure 1 suggests to define Regions I, II and III. In Region II the sign test is more powerful than Fisher's exact test; in Regions I and III the exact test is more powerful.

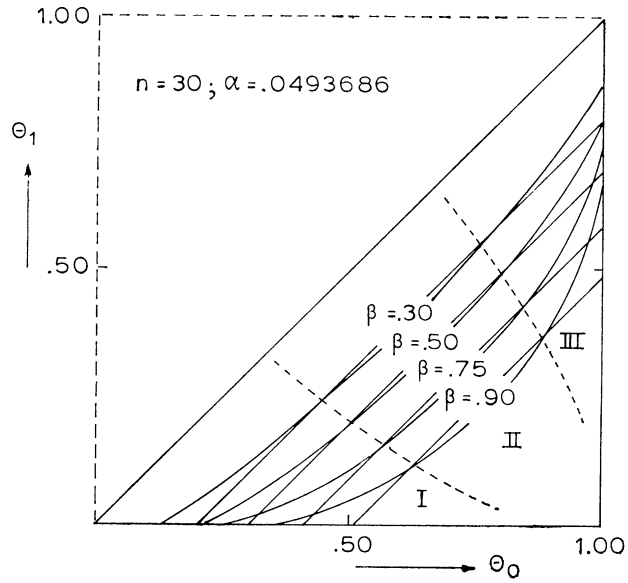

FIG. 1.

The following notion

$$\gamma(\varphi, \varphi') = \sup_{\theta \in \Theta_1} \{E_\theta(\varphi') - E_\theta(\varphi)\} \tag{5.1}$$

of the maximum shortcoming of test φ with respect to test φ' seems to be useful, though often misleading. Figure 1 shows that φ_{hyp} is only a little bit less powerful than φ_{sign} in Region II (the computations gave $\gamma(\varphi_{\text{hyp}}, \varphi_{\text{sign}}) = .038$) whereas Regions I and III contain points where φ_{sign} is much less powerful than φ_{hyp} (we found $\gamma(\varphi_{\text{sign}}, \varphi_{\text{hyp}}) \approx .22$). The latter points however are not very likely to occur because the experimenter will usually expect that the order-effect is not

extremely large or, in formula, that θ will not be far from the line $\theta_0 + \theta_1 = 1$ which corresponds to no order-effects.

We made similar comparisons for $n = 10, 15, 20, 25, 30, 37, 46$, and 50 with $\alpha \approx .05$ such that $\gamma_{n,\alpha} = 1$. For $n = 10$ we did not find any point $\theta \in \Theta_1$ with $E_\theta(\varphi_{\text{hyp}}) > E_\theta(\varphi_{\text{sign}})$. This indicates that φ_{hyp} might be inadmissible, at least sometimes (see Section 7). As n increases, Region II becomes smaller and smaller (see Figure 2) just like $\gamma(\varphi_{\text{hyp}}, \varphi_{\text{sign}})$ (see Figure 3), while $\gamma(\varphi_{\text{sign}}, \varphi_{\text{hyp}})$ becomes larger and larger (Figure 4). This indicates that, as *n increases, Fisher's exact test becomes more and more attractive whereas the sign test becomes less attractive.*

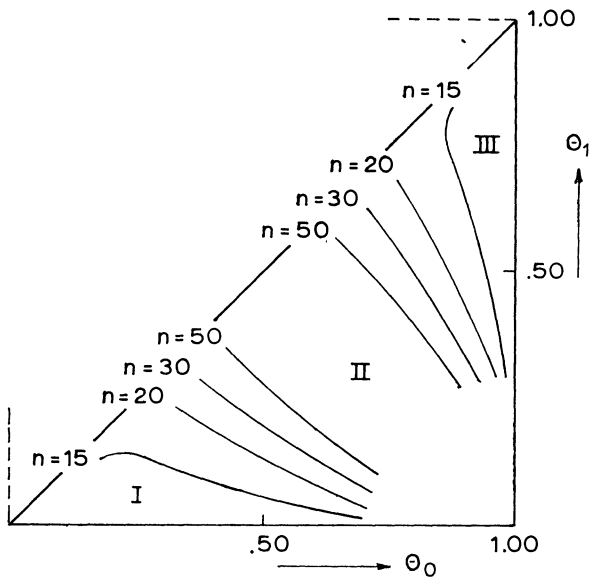

FIG. 2.

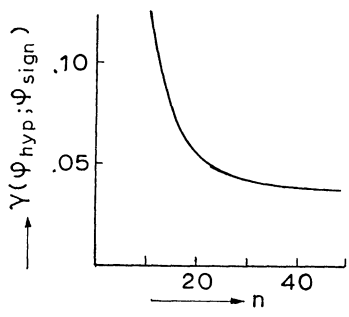

FIG. 3.

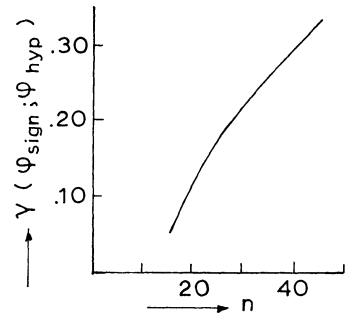

FIG. 4.

6. The class of all SMP unbiased size-α tests. Let E denote an arbitrary class of tests. A test φ' is said to be SMP (E) (*somewhere most powerful among the tests of class E*) if (i) $\varphi' \in E$ and (ii) for at least one point $\theta \in \Theta_1$ we have $E_\theta(\varphi') = \sup_{\varphi \in E} E_\theta(\varphi)$. We call φ' SMP level-α if E is the class of all level-α tests; if E

denotes the class of all unbiased size-α tests, then φ' is said to be SMP unbiased size-α (see [5]).

By considering simple alternatives of the form $\theta = (\frac{1}{2} + \rho, \frac{1}{2} - \rho)\,(0 < \rho < \frac{1}{2})$ corresponding to "no order-effects," the sign test can be shown to be both SMP level-α (consider $\theta_0 = (\frac{1}{2}, \frac{1}{2})$ as the corresponding least favorable parameter in Θ_0) and SMP unbiased size-α. No level-α test can improve upon the sign test if θ is of the above-mentioned form: Region II of Section 5 cannot be empty, Fisher's exact test will not be UMP unbiased. It follows easily that the sign test is uniformly maximin in the following sense: φ_{sign} is maximin level-α for testing H against the subalternative A_ν: $\theta_0 \geqq \theta_1 + \nu$ for each value of ν $(0 < \nu < 1)$. This, however, is not a very compelling optimum property.

We deepen the insight into our problem (H, A) by constructing *the class C of all* SMP *unbiased size-α tests*; this will be done in the usual way via similarity and Neyman-structure. Accordingly, let D_u denote the class of all unbiased size-α tests for our testing problem (H, A), let D_s denote the class of all similar size-α tests for testing Hypothesis H and let D_n denote the class of all tests having Neyman-structure.

REMARK. One would like to apply the standard theory of multiparameter exponential families (Lehmann [3] Section 4.4, Ferguson [2] Section 5.4). Difficulties appear "because $(S_{00}S_{01}S_{10})$ is 3-dimensional whereas $\theta = (\theta_0, \theta_1)$ is 2-dimensional." We shall avoid these difficulties by working with the basic concepts.

First we prove $D_u \subset D_s = D_n$ and we characterize D_n. For that purpose consider $\{P_\theta;\ \theta \in \Theta_0\}$. It is shown easily that $Y = S_{00} + S_{10}$ is a *complete* sufficient statistic for this family. Thus (see Lehmann or Ferguson) $D_s = D_n$ and $\varphi \in D_n$ if and only if

$$E_\theta\{\varphi(S_{00}S_{01}S_{10}S_{11}) \mid Y = y\} = \alpha \tag{6.1}$$

for all $\theta \in \Theta_0$ and $y = 0, 1, \cdots, n$.

We need the conditional distribution of $(S_{00}, \cdots, S_{11})$ given $Y = y$. It follows from Section 3 that Y has the binomial $B\{n;\ 1 - \frac{1}{2}(\theta_0 + \theta_1)\}$ distribution. Computing the conditional probability that $(S_{00}, \cdots, S_{11}) = (s_{00}, \cdots, s_{11})$ given $Y = y$ for arbitrary $\theta \in \Theta$, we find that S_{01} *and* S_{10} *are conditionally independent*, S_{01} *having the binomial* $B\{n - y;\ \theta_0/(\theta_0 + \theta_1)\}$ *distribution and* S_{10} *having the binomial* $B\{y;\ (1 - \theta_1)/(2 - \theta_0 - \theta_1)\}$ *distribution*. Thus

$$\begin{aligned} P_\theta\{(S_{00} = y - s_{10}, S_{01} = s_{01}, S_{10} = s_{10}, S_{11} = n - y - s_{01}) \mid Y = y\} \\ = P_\theta(S_{01} = s_{01} \mid Y = y)P_\theta(S_{10} = s_{10} \mid Y = y) \end{aligned} \tag{6.2}$$

where the right-hand side is a product of two binomial probabilities. For $\theta \in \Theta_0$ the right-hand side of (6.2) becomes $\binom{n-y}{s_{01}}\binom{y}{s_{10}}2^{-n}$ independently of θ, as it should be because Y is sufficient for the family $\{P_\theta;\ \theta \in \Theta_0\}$.

Let $\theta = (\theta_0, \theta_1)$ be an arbitrary point in the interior of Θ_1 (there arise some unpleasant and uninteresting degeneracies if $\theta = (\rho, 0)$ or $(1, \rho)$; we restrict our attention to the interior of Θ_1). We can maximize $E_\theta(\varphi)$ among all tests with

Neyman-structure by applying the Neyman–Pearson Fundamental Lemma to the conditional distributions given $Y = y$. By doing so, we find that we have to reject for large values of the statistic

$$T_\rho = \rho S_{01} + (1 - \rho)S_{10} \tag{6.3}$$

where $\rho \in (0, 1)$ is defined by

$$\rho = \log (\theta_0/\theta_1)/\{\log (\theta_0/\theta_1) + \log (1 - \theta_1)/(1 - \theta_0)\} . \tag{6.4}$$

Under the hypothesis, the distribution P_0 of T_ρ given $Y = y$, is that of a weighted sum (6.3) of two independent binomial variables, S_{01} having the $B(n - y; \frac{1}{2})$ and S_{10} the $B(y; \frac{1}{2})$ distribution. Let $t_{n,y,\rho,\alpha}$ be determined such that it is the smallest real number t satisfying

$$P_0(T_\rho > t \mid Y = y) < \alpha \tag{6.5}$$

and define

$$\gamma_{n,y,\rho,\alpha} = \{\alpha - P_0(T_\rho > t_{n,y,\rho,\alpha} \mid Y = y)\}/P_0(T_\rho = t_{n,y,\rho,\alpha} \mid Y = y) . \tag{6.6}$$

We can now formulate the basic theorem which characterizes the class C of all SMP unbiased size-α tests by describing for arbitrary fixed θ in the interior of Θ_1 the tests φ^* that are MP in θ among the unbiased size-α tests for Problem (H, A).

In Section 3 we defined tests as functions over the outcome space of $(S_{00} S_{01} S_{10} S_{11})$. There exists a 1 : 1 correspondence between this outcome space and the range of the random vector $(S_{01} S_{10} Y)$. Thus we can define our tests equally well as functions over the outcome space $(S_{01} S_{10} Y)$. This is a bit more convenient.

THEOREM. *Necessary and sufficient for* $\varphi^* \in D_u$, $E_\theta(\varphi^*) = \sup_{\varphi \in D_u} E_\theta(\varphi)$ *is that for this fixed value of* θ, ρ *is determined by* (6.4),

$$\begin{aligned} \varphi^*(s_{01} s_{10} y) &= 0 \quad \text{if} \quad \rho s_{01} + (1 - \rho)s_{10} < t_{n,y,\rho,\alpha} \\ &= 1 \quad \text{if} \quad \rho s_{01} + (1 - \rho)s_{10} > t_{n,y,\rho,\alpha} \end{aligned} \tag{6.7}$$

while in the remaining points φ^* *has to be determined such that* (6.1) *holds, or equivalently*

$$\sum \varphi^*(s_{01} s_{10} y)\binom{n-y}{s_{01}}\binom{y}{s_{10}}2^{-n} = \alpha - P_0(T_\rho > t_{n,y,\rho,\alpha} \mid Y = y) \tag{6.8}$$

where the summation is taken over all (s_{01}, s_{10}) *satisfying* $\rho s_{01} + (1 - \rho)s_{10} = t_{n,y,\rho,\alpha}$.

An interesting test φ^* *satisfying* (6.7) *and* (6.8) *is obtained by applying "uniform randomization," that means by defining* $\varphi^*(s_{01} s_{10} y)$ *equal to the constant* $\gamma_{n,y,\rho,\alpha}$ *for all* (s_{01}, s_{10}) *satisfying* $\rho s_{01} + (1 - \rho)s_{10} = t_{n,y,\rho,\alpha}$. *The test obtained is called* φ_ρ.

PROOF. The discussion preceding the formulation of the theorem contains a proof for the theorem if the class $D_s = D_n$ is considered, instead of the class D_u which is a subclass of D_s. Our proof is complete if we can show that each test φ^* satisfying (6.7) and (6.8), automatically belongs to the class D_u of all unbiased

size-α tests for Problem (H, A). It follows from the construction of φ^* and particularly from (6.8) that φ^* has Neyman-structure and that φ^* is similar size-α. Hence, a sufficient condition for the unbiasedness of φ^* is that $E_\theta(\varphi^* \mid Y = y) \geqq \alpha$ for all y and all $\theta \in \Theta_1$. This will be proved in the following part of this section ending at "interpretation of the theorem" (this part is not essential for a good understanding of the rest of the paper).

It follows from the basic probabilistic result above (6.2) that both S_{01} and S_{10} are stochastically larger if $\theta \in \Theta_1$ than for $\theta' \in \Theta_0$, when conditioning on $Y = y$. S_{01} and S_{10} being conditionally independent, one immediately infers that $T_\rho = \rho S_{01} + (1 - \rho)S_{10}$ where $\rho \in (0, 1)$ is stochastically larger if $\theta \in \Theta_1$ than for $\theta' \in \Theta_0$. Thus for $\theta \in \Theta_1$ and $\theta' \in \Theta_0$, we have

$$\alpha = E_{\theta'}(\varphi_\rho \mid Y = y) \leqq E_\theta(\varphi_\rho \mid Y = y)$$

because φ_ρ is a non-decreasing function of T_ρ. This proves the unbiasedness of test φ_ρ which uses "uniform randomization."

For almost all $\rho \in (0, 1)$, φ_ρ will be the unique test satisfying (6.7) and (6.8) (see interpretations of the theorem). But for some rational values of ρ and particularly for the very important value $\rho = \frac{1}{2}$ (see Section 7), φ_ρ will not be the only test φ^* satisfying (6.7) and (6.8). How to prove the unbiasedness of φ^* in these cases? Remark that φ^* will not be a function of T_ρ: the value of φ^* is not uniquely determined if $\rho s_{01} + (1 - \rho)s_{10} = t_{n,y,\rho,\alpha}$, at least for some values of y. In order to deal with these cases, we extended the above-mentioned arguments as follows. Remark that, when condioning on $Y = y$: (i) S_{01} and S_{10} are independent, (ii) S_{01} and S_{10} are stochastically larger if $\theta \in \Theta_1$ than for $\theta' \in \Theta_0$, (iii) if $s'_{01} \geqq s_{01}$ and $s'_{10}, \geqq s_{10}$, then $\varphi^*(s_{01}\, s_{10}\, y) \leqq \varphi^*(s'_{01}\, s'_{10}\, y)$. The lemma below provides that for each φ^* satisfying (6.7) and (6.8),

$$\alpha = E_{\theta'}\{\varphi^*(S_{01}\, S_{10}\, y) \mid Y = y\} \leqq E_\theta(\varphi^* \mid Y = y)\,.$$

This completes the proof for the unbiasedness of φ^*.

Let $\leqq$ denote the natural partial ordering in R^p: we say $x \leqq y$ iff $x_i \leqq y_i$ $(i = 1, \cdots, p)$. A function $\varphi : R^p \to R$ is said to be non-decreasing w.r. to $\leqq$ if $x \leqq y$ implies $\varphi(x) \leqq \varphi(y)$ (φ is "isotonic" in the sense of Barlow, Bartholomew, Bremmer and Brunk).

LEMMA. *Let $X = (X_1, \cdots, X_p)$ and $Y = (Y_1, \cdots, Y_p)$ be two p-variate random vectors and let $\varphi : R^p \to R$ be non-decreasing* w.r. *to* $\leqq$. *If* (i) $X_1, \cdots, X_p$ *are independent*, (ii) $Y_1, \cdots, Y_p$ *are independent and* (iii) Y_i *is stochastically larger than* X_i $(i = 1, \cdots, p)$, *then* $\varphi(Y)$ *is stochastically larger than* $\varphi(X)$ *with as a result that* $E\{\varphi(Y)\} \geqq E\{\varphi(X)\}$ *if the expectations exist.*

PROOF. Assume that X_i and Y_i have *continuous* distribution functions F_i and G_i respectively, which are *strictly increasing* on the same interval $I_i = (a_i, b_i)$ and constantly 0 or 1 outside of this interval ($a_i = -\infty$ and $b_i = \infty$ permitted; $i = 1, \cdots, p$). Define $f_i : I_i \to I_i$ by $f_i(x) = F_i^{-1}\{G_i(x)\}$. Observe that X_i and

$f_i(Y_i)$ have the same distribution while $f_i(y_i) \leqq y_i$ for all $y_i \in I_i$. The independence assumptions in the lemma imply that $\varphi(X) = \varphi(X_1, \cdots, X_p)$ and $\varphi\{f_1(Y_1), \cdots, f_p(Y_p)\}$ have the same distribution. Hence

$$P\{\varphi(X) \leqq z\} = P[\varphi\{f_1(Y_1), \cdots, f_p(Y_p)\} \leqq z] \geqq P\{\varphi(Y) \leqq z\}$$

because $\varphi(y_1 \cdots y_p) \leqq z$ implies $\varphi\{f_1(y_1), \cdots, f_p(y_p)\} \leqq z$. See Remark 3 for further discussions.

REMARK 1. Dealing with φ_ρ in the proof of the theorem, we remarked that one immediately infers that T is stochastically larger for $\theta \in \Theta_1$ than for $\theta' \in \Theta_0$. Though this result can be proved easily, it is interesting to remark that it can also be derived from the lemma: let T play the part of φ.

REMARK 2. It might be interesting to define that a p-variate rv Y is stochastically larger than a p-variate rv X if $\varphi(Y)$ is stochastically larger than $\varphi(X)$ for each $\varphi: R^p \to R$ which is non-decreasing w.r. $\leqq$. Necessary and sufficient condition for $X \leqq Y$ in this sense is that $\varphi(Y)$ is stochastically larger than $\varphi(X)$ for each non-decreasing *indicator* function $\varphi = I_A$. The lemma provides sufficient conditions for $X \leqq Y$ based on independence assumptions. A necessary but insufficient condition for $X \leqq Y$ is that $F \geqq G$ holds for the p-dimensional distribution functions F and G of X and Y respectively.

REMARK 3. The assumptions used in the proof of the lemma should be removed. This can be done by defining $F_i^{-1}(u) = \inf\{x; F_i(x) \geqq u\}$. If G_i is continuous, then all arguments in the proof of the lemma hold. If G_i is not continuous then $f_i(y_i) \leqq y_i$ holds, but X_i and $f_i(Y_i)$ do not necessarily have the same distribution. In fact there exists a function ψ_i such that $\psi_i(x) \geqq x$ and $\psi_i(X_i)$ and $f_i(Y_i)$ have the same distribution. Now the proof of the lemma can be completed by extending the arguments used at the end.

Interpretations of the theorem. We are interested in (i) when is φ_ρ the unique test satisfying (6.7) and (6.8), (ii) what does the class C of all SMP unbiased size-α tests look like? With respect to (i) we remark that uniqueness will hold for all irrational and nearly all rational values of ρ because there will usually exist exactly one point (s_{01}, s_{10}) satisfying $\rho s_{01} + (1 - \rho)s_{10} = t_{n,y,\rho,\alpha}$ (whatever y may be). Uniqueness will also hold if, for the values of y of interest, φ_ρ is essentially nonrandomized; that means if α is such that (6.8) can only be satisfied by taking $\varphi^*(s_{01}\, s_{10}\, y) = 1$. With respect to (ii) we remark that different values of ρ may also provide the same test satisfying (6.7) and (6.8). The situation will be as follows. Let $\rho(\theta)$ be defined by (6.4). There will exist rational numbers $0 < \rho_1 < \cdots < \rho_k < \frac{1}{2} < 1 - \rho_k < \cdots < 1 - \rho_1 < 1$, partitioning $[0, 1]$ into $2k + 2$ subintervals, such that all θ with $\rho(\theta)$ in the interior of one of these subintervals provide the same unique optimum test satisfying (6.7) and (6.8), while for θ's with $\rho(\theta)$ equal to one of the separating points the corresponding optimum test is not uniquely determined because randomization need not be uniform.

If $\alpha \leqq 2^{-n}$, then no separating points exist: for each value of ρ we obtain the same test satisfying (6.7) and (6.8). This test rejects with probability $\alpha \cdot 2^{-n}$ if $(s_{01}\, s_{10}) = (n - y, y)$, and accepts otherwise. Thus it is the sign test: *the sign test is the unique* UMP *unbiased size-α test if* $\alpha \leqq 2^{-n}$. Similar results hold for $\alpha \geqq 1 - 2^{-n}$.

For $2^{-n} < \alpha < 1 - 2^{-n}$ there does not exist a UMP unbiased size-α test. The basic theorem determines the class C of all SMP unbiased size-α tests. It is verified easily that the sign test and $\varphi_{\frac{1}{2}}$ are equivalent: $\varphi_{\frac{1}{2}}(S_{01}\, S_{10}\, Y)$ and $\varphi_{\text{sign}}(S_{00}\, S_{01}\, S_{10}\, S_{11})$ define the same test-function over $\mathscr{X}$, or to put it otherwise

$$\varphi_{\text{sign}}(s_{00}\, s_{01}\, s_{10}\, s_{11}) = \varphi_{\frac{1}{2}}(s_{01},\, s_{10},\, s_{00} + s_{10})$$

7. Admissibility and invariance. The sign test is admissible if $\alpha \leqq 2^{-n}$ or $\alpha \geqq 1 - 2^{-n}$, because it is then the unique UMP unbiased size-α test. For $2^{-n} < \alpha < 1 - 2^{-n}$ we have to distinguish two cases: (i) the sign test is actually nonrandomized; $\gamma_{n,\alpha} = 1$ in (4.1) and (4.2); (ii) the sign test is genuinely randomized; $0 < \gamma_{n,\alpha} < 1$. In Case (i) the sign test is admissible because it is the *unique* test maximizing the power in $\theta = (\frac{1}{2} + \nu, \frac{1}{2} - \nu)$ $(0 < \nu < \frac{1}{2})$.

LEMMA. *The sign test is inadmissible if* $2^{-n} < \alpha < 1 - 2^{-n}$; $0 < \gamma_{n,\alpha} < 1$ *and* $n \geqq 3$.

PROOF. We shall construct a test φ_{mod} with $E_\theta(\varphi_{\text{mod}}) > E_\theta(\varphi_{\text{sign}})$ for all $\theta \in \Theta_1$ with $\rho(\theta) \neq \frac{1}{2}$ and equality of the power functions for $\rho(\theta) = \frac{1}{2}$. The idea is as follows. φ_{sign} is equivalent to $\varphi_{\frac{1}{2}}$, which test applies uniform randomization to (6.7), (6.8) in case $\rho = \frac{1}{2}$. φ_{mod} has to satisfy (6.7), (6.8) for $\rho = \frac{1}{2}$. The idea is that uniform randomization is worse than some other way. Remark that Y has the binomial $B(n, 1 - \frac{1}{2}\theta_0 - \frac{1}{2}\theta_1)$ distribution. Thus Y contains information with respect to the direction of the order-effect: whether $\theta_0 + \theta_1 > 1$ (or equivalently $\rho(\theta) < \frac{1}{2}$, see (6.4)) or $\theta_0 + \theta_1 < 1$ (or equivalently $\rho(\theta) > \frac{1}{2}$). A small outcome y of Y suggests to work with $\rho < \frac{1}{2}$ and to reject first large values of s_{10}. These considerations suggest to define φ_{mod} as follows. (In order to get an easy proof of the superiority of φ_{mod} over $\varphi_{\text{sign}} \sim \varphi_{\frac{1}{2}}$, we modify $\varphi_{\frac{1}{2}}$ only slightly, that means only for $y = 1$ and $y = n - 1$.)

We define $\varphi_{\text{mod}}(s_{01}\, s_{10}\, y) = \varphi_{\frac{1}{2}}(s_{01}\, s_{10}\, y)$, unless $[y = 1$ and $\varphi_{\frac{1}{2}}(s_{01},\, s_{10},\, 1) = \gamma_{n,\alpha} = \gamma_{n,1,\frac{1}{2},\alpha}$ (see (4.1) and (6.6))], or $[y = n - 1$ and $\varphi_{\frac{1}{2}}(s_{01},\, s_{10},\, n - 1) = \gamma_{n,\alpha} = \gamma_{n,n-1,\frac{1}{2},\alpha}]$. In the remaining four points we define $\varphi_{\text{mod}}(s_{01}\, s_{10}\, 1)$ as large as possible for the point with $s_{10} = 1$, and $\varphi_{\text{mod}}(s_{01},\, s_{10},\, n - 1)$ as large as possible for $s_{01} = 1$. One can compute the difference in power between φ_{mod} and φ_{sign} by restricting attention to the four points where they differ. Along these lines we found that this power difference is always nonnegative with the result that the sign test is inadmissible.

What about the admissibility of Fisher's exact test? This test does not belong to the class C of SMP unbiased size-α tests (verify the case $n = 1$ for example). It follows that φ_{hyp} is inadmissible in case $\alpha \leqq 2^{-n}$ and $\alpha \leqq 1 - 2^{-n}$ because

then φ_{sign} is the unique UMP unbiased size-α test. The computations showed for some other situations ($n \leqq 10$, $\alpha = .05$) that φ_{sign} is uniformly better than φ_{hyp} with the result that φ_{hyp} will then be inadmissible. But usually neither the sign test nor the exact test is uniformly better than the other. We have no proof for the inadmissibility of φ_{hyp} in these cases.

We have seen that usually ($2^{-n} < \alpha < 1 - 2^{-n}$) no UMP unbiased size-$\alpha$ test exists. One would like to apply invariance considerations. Consider the problem which is obtained after reduction by sufficiency (Section 3). The only group G leaving the testing problem invariant (and which we could invent) consists of two bijections e and $g = g^{-1}$ where e is the identical mapping and

$$g(s_{00}\, s_{01}\, s_{10}\, s_{11}) = (s_{11}\, s_{10}\, s_{01}\, s_{00}) .$$

Using the notation of Ferguson [2], we find that the induced mapping $\bar{g}: \Theta \to \Theta$ is defined by

$$\bar{g}(\theta_0, \theta_1) = (1 - \theta_1, 1 - \theta_0) .$$

This obviously leaves the testing problem invariant. The tests φ_{sign}, φ_{hyp}, $\varphi_{\frac{1}{2}}$ and φ_{mod} are invariant. This shows that there usually will not exist a UMP test, even not when attention is restricted to tests that are both unbiased and invariant.

8. Other designs for counterbalancing the order effects. The preceding sections were based on the "coin-tossing" design where drug and placebo are administered to the ith patient in an order which is determined by tossing (independently) a fair coin. Some experimenters did not take all precautions that belong to this design; or they simply preferred another sampling scheme. We shall discuss three alternative designs: (i) the "unfair-coin" design, (ii) the "alternating" design, (iii) the "inexplicable" design. Of course each design leads to a new probabilistic model and accordingly to a new theory.

Sometimes the order is determined by means of a random mechanism which is not manipulated by the experimenter and which simulates independent tossings with an unfair coin. Suppose one wants to "prove" the supposition that "in a cat the response to electric stimulation in the cortex is more vehement in the drowsy state than in the awake state." The null-hypothesis should be tested that "in a cat the response does not depend upon the initial behavioral state." It has to be remarked that the meaning of the conclusions to be obtained is always restricted by the means of experimentation. Suppose our experimenter, aiming at an efficient design, stimulated each of n cats twice. The ith cat was stimulated first in the state which presented itself first ($x_{2i-1} = 0$ if this was the awake state and $x_{2i-1} = 1$ if it was the drowsy one). Next he waited till the cat arrived in the other state, he stimulated again and denoted whether the response was more vehement in the first state ($x_{2i} = 0$) or in the second ($x_{2i} = 1$). [In fact one of my clients used a procedure in which he distinguished seven initial behavioral states which were stimulated in the order in which they presented themselves to the observer; it would have consumed too much time if a predetermined random order had been used.] First we develop a probabilistic model

for this "unfair-coin" design. Assume that the n random vectors (X_{2i-1}, X_{2i}) are independently and identically distributed (all cats were treated by means of the same stimulation procedure). The joint distribution of (X_{2i-1}, X_{2i}) can be determined by means of the following three parameters: $\pi = P(X_{2i-1} = 1)$ denoting the probability that a "randomly drawn" cat presents the drowsy state first, and θ_j $(j = 0, 1)$ according to (2.1). Remark that (X_{2i-1}, X_{2i}) assumes the outcomes $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$ with probabilities $(1 - \pi)(1 - \theta_0)$, $(1 - \pi)\theta_0$, $\pi(1 - \theta_1)$ and $\pi\theta_1$ respectively. Thus P_θ belonging to parameter $\theta = (\pi, \theta_0, \theta_1)$ is determined by

$$(8.1) \qquad P_\theta(\{x\}) = (1 - \pi)^{Z(x)}\pi^{n-Z(x)}(1 - \theta_0)^{S_{00}(x)}\theta_0^{S_{01}(x)}(1 - \theta_1)^{S_{10}(x)}\theta_1^{S_{11}(x)}$$

where we use the notation (3.1) (see also (2.2) and (3.2)). Next we try to express null-hypothesis and alternative in terms of our parameter θ. The referee observed that there exists some ambiguity here. Remark that "in a cat, and experimenting along the lines described, the probability to get the highest response in the drowsy state," in formula $P_\theta\{X_{2i-1} \neq X_{2i}\}$, is equal to

$$\gamma = (1 - \pi)\theta_0 + \pi(1 - \theta_1) = \tfrac{1}{2} + \tfrac{1}{2}(\theta_0 - \theta_1) + (\pi - \tfrac{1}{2})(1 - \theta_0 - \theta_1)\,.$$

In Sections 2—7 we assumed $\pi = \frac{1}{2}$ and testing $H\colon \theta_0 = \theta_1$ against $A\colon \theta_0 > \theta_1$ is in complete agreement with testing $H'\colon \gamma = \frac{1}{2}$ against $A'\colon \gamma > \frac{1}{2}$. But for the "unfair-coin" design the testing problems (H, A) and (H', A') are different. Even the probabilistic models differ because for (H, A) it is assumed that $\theta_0 \geqq \theta_1$ whereas for (H', A') it is assumed that $\gamma \geqq \frac{1}{2}$. Neither of the corresponding regions in $\theta = (\pi, \theta_0, \theta_1)$-space is a subset of the other. Both (H, A) and (H', A') are solved easily by constructing UMP similar size-α tests and applying the usual theory of multiparameter exponential families starting from the observation that $(S_{00}S_{01}S_{10}S_{11})$ is a complete sufficient statistic having the multinomial $M\{n; (1 - \pi)(1 - \theta_0), (1 - \pi)\theta_0, \pi(1 - \theta_1), \pi\theta_1\}$ distribution if $\theta = (\pi, \theta_0, \theta_1)$. Fisher's exact test (4.5) is UMP unbiased size-α for problem (H, A); the sign test (4.2) is UMP unbiased size-α for problem (H', A'). One verifies immediately that Fisher's exact test is not a level-α test for testing Hypothesis H', and the sign test is not a level-α test for testing H. We have the opinion that usually the appropriate formulation is to test $H\colon \theta_0 = \theta_1$ against $A\colon \theta_0 > \theta_1$ with as a consequence that Fisher's test should, and the sign test should not, be applied to the case of an "unfair-coin" design. First remark that the original formulation "in a cat the response does not depend upon the initial behavioral state" might very well be translated into H: "the conditional probability to get the highest response in the state tried last does not depend on whether this is the awake state or the drowsy one." Next remark that the original formulation need not be in agreement with $H'\colon \gamma = \frac{1}{2}$. For that purpose suppose that (i) there exists a bias in favor of the second stimulation ("overreaction," $\theta_0 + \theta_1 > 1$), (ii) the response to the first stimultation does not depend on whether this happens in the awake or in the drowsy state; the same holds for

the response in the second stimulation (though this tends to be larger according to (i)), (iii) the cat usually presents itself first in the awake state ($\pi < \frac{1}{2}$). It follows that $\gamma > \frac{1}{2}$: it looks as if drowsiness facilitates high scores, but this is only due to overreaction and the fact that the drowsy state is usually the last one. In the rest of this paper we shall consider some other designs, but always restrict the attention to testing $H\colon \theta_0 = \theta_1$ against $A\colon \theta_0 > \theta_1$.

Many experimenters feel reluctant to apply the coin-tossing design because they do not want to introduce "more randomness." They prefer a design where $x_1, x_3, \cdots, x_{2n-1}$ is a predetermined sequence of 0's and 1's. Usually they use the "alternating" design where $x_{2i-1} = 0$ if i is odd and $x_{2i-1} = 1$ if i is even ($i = 1, \cdots, n$). Now the sample space $\mathcal{X}$ essentially consists of all 2^n points $(x_2\, x_4 \cdots x_{2n})$ where $x_{2i} = 0$ or 1. What we get is the classical problem for comparing two probabilities. One can complete the 2×2-table of Section 3. *Fisher's exact test* (4.5) *is* UMP *unbiased size-*α. For a continuation see Section 9.

Some experimenters determined $x_1, x_3, \cdots, x_{2n-1}$ in an "inexplicable" way, introducing an *interdependence* between the rv's $X_1, X_3, \cdots, X_{2n-1}$ which is difficult to describe. As an example, consider the experiment with n cats, described at the beginning of this section. During the experiment the experimenter might start to believe that most cats present the awake state first ("$\pi < \frac{1}{2}$") and that it is the task of the experimenter to do something about it. He might start ignoring some awake states in order to get more cats presenting the drowsy state first. Other experimenters are mentally unable to toss coins independently. If they find six heads in a row or if they otherwise obtained too many heads in the past, then they cannot resist the temptation to ignore some of these results by tossing again.

A safe model for this "inexplicable" design is as follows. We do not make any probabilistic assumption concerning the joint distribution of $X_1, X_3, \cdots, X_{2n-1}$. We only assume some kind of conditional independence such that we may write

$$P_\theta(\{x\}) = P\{X_{2i-1} = x_{2i-1} (i = 1, \cdots, n)\} \times \prod_{i=1}^{n} P_\theta(X_{2i} = x_{2i} \mid X_{2i-1} = x_{2i-1}) \tag{8.2}$$

where θ is of the form $\theta = (\theta_0, \theta_1, P)$, θ_0 and θ_1 being defined in (2.1) and P denoting an arbitrary probability distribution over the space of all 2^n sequences of n numbers zero or one. The second factor in the right-hand side of (8.2) is a product consisting of factors θ_0, θ_1, $1 - \theta_0$, $1 - \theta_1$. We again have to test $H\colon \theta_0 = \theta_1$ against $A\colon \theta_0 > \theta_1$. The parameter space Θ is pretty intricate. One might write formally $\Theta = \Delta \times \mathcal{P}$ where Δ denotes the set of all (θ_0, θ_1)'s satisfying $0 \leqq \theta_1 \leqq \theta_0 \leqq 1$ and where $\mathcal{P}$ denotes the $(2^n - 1)$-dimensional set of all possible probability distributions P over the space of all 2^n sequences of n numbers zero or one; P is determined by giving the probabilities of the elementary events (except one) in the latter outcome space. Of course we cannot apply the usual theory for multiparameter exponential families. In Section 10 we shall nevertheless show that *Fisher's exact test* (4.5) *is again* UMP *unbiased size-*α. It is obvious that the sign test is not a level-α test here (see the "unfair-coin" design).

9. The alternating design if n is even. We shall show that the sign test is a reasonable competitor of Fisher's UMP unbiased exact test and we shall investigate whether the alternating design is indeed better than the coin-tossing design "because less randomness is introduced." We assume that n is even.

We have met experimenters who used the alternating design and next applied the sign test. We can show that the sign test is a level-α test if $\alpha \leqq \frac{1}{2}$.

LEMMA. *If $\theta = (\rho, \rho) \in \Theta_0$ and $b \geqq \frac{1}{2}n$, then $P_\theta(S \geqq b)$ is a strictly increasing function of $\pi = \rho(1 - \rho)$ with its maximum equal to $P_{(\frac{1}{2},\frac{1}{2})}(S \geqq b) = P(B \geqq b)$ where B has the binomial $B(n, \frac{1}{2})$-distribution. Hence $E_\theta(\varphi_{\text{sign}}) \leqq \alpha$ for all $\theta \in \Theta_0$ if $\alpha \leqq \frac{1}{2}$.*

PROOF. If $\theta = (\rho, \rho)$ then S_{01} has the binomial $B(\frac{1}{2}n, \rho)$ and S_{10} the binomial $B(\frac{1}{2}n, 1 - \rho)$ distribution. Hence $S = S_{01} + S_{10}$ has the same distribution as $\frac{1}{2}n + D_1 + \cdots + D_{\frac{1}{2}n}$ where the D_i's are i.i.d. with $P_\theta(D_i = -1) = P_\theta(D_i = 1) = \pi$ and $P_\theta(D_i = 0) = 1 - 2\pi$. Remark that $\pi = \rho(1 - \rho) \leqq 4^{-1}$. We can prove that $P_\pi(D_1 + \cdots + D_m \geqq c)$ is a strictly increasing function of π if $c > 0$ and $\pi \leqq \frac{1}{3}$. For $\pi > \frac{1}{3}$ the distribution of D_i is no longer unimodal and one can construct counter-examples by considering $m = 2$.

The lemma shows that the sign test is a level-α test but that it will not be unbiased size-α because $E_\theta(\varphi_{\text{sign}}) < \alpha$ for $\theta = (\rho, \rho)$ with $\rho \neq \frac{1}{2}$ and $E_\theta(\varphi_{\text{sign}})$ continuous in θ. Hence Fisher's exact test need not be uniformly more powerful. In fact it can be shown that no level-α test can be more powerful than the sign test for any point on the line-segment between $(\frac{1}{2}, \frac{1}{2})$ and $(1, 0)$: there the sign test is uniformly best among all tests satisfying $E_{(\frac{1}{2},\frac{1}{2})}(\varphi) \leqq \alpha$.

We compared φ_{sign} and φ_{hyp} numerically for $n = 10$, 20, 30, and 46 with

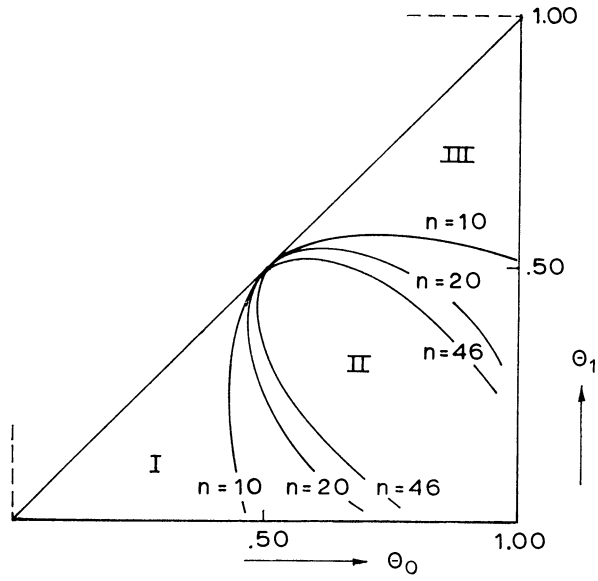

FIG. 5.

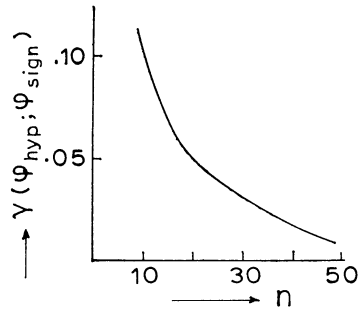

FIG. 6.

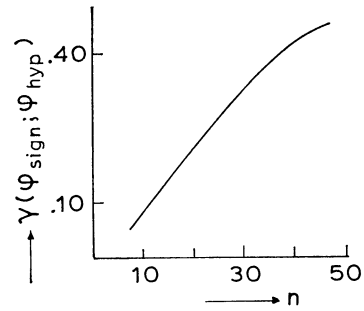

FIG. 7.

$\alpha \approx .05$ such that φ_{sign} is non-randomized. The results are summarized in Figures 5, 6 and 7 (for an explanation, see Section 5). As n increases, Region II (this is the set of all $\theta = (\theta_0, \theta_1)$'s with $E_\theta(\varphi_{\text{sign}}) > E_\theta(\varphi_{\text{hyp}})$) becomes smaller and smaller (see Figure 5) just like $\gamma(\varphi_{\text{hyp}}, \varphi_{\text{sign}})$ (see Figure 6) while $\gamma(\varphi_{\text{sign}}, \varphi_{\text{hyp}})$ becomes larger and larger (see Figure 7). This indicates that as n increases, φ_{hyp} becomes more attractive, whereas the sign test gets less attractive. The interesting point in this comparison is that for small values of n the sign test is more attractive from an over-all point of view than Fisher's UMP unbiased exact size-α test φ_{hyp}: the unbiasedness restriction is too restrictive in case n is small.

This result has interpretations for the usual problem of testing whether two success-probabilities are different, on the basis of two independent series of Bernoulli trials. If both series consist of $\frac{1}{2}n$ trials and n is small (say $n = 10$), then we prefer to use the sign test instead of Fisher's exact test. This sign test is carried out by testing $s_1 + f_2$ in the $B(n, \frac{1}{2})$ distribution where s_1 is the number of successes in the first series and f_2 is the number of failures in the second one.

Finally we want to compare the coin-tossing design and the alternating design. We have conjectured that the latter design would be better in the following sense. If φ is an arbitrary (unbiased) test for the testing problem based on the coin-tossing design (Sections 2—7), then there will exist a better test φ' for the testing problem based on the alternating design (and the same sample size n of course): $E_\theta(\varphi') \leqq E_\theta(\varphi)$ for all $\theta \in \Theta_0$ and $E_\theta(\varphi') \geqq E_\theta(\varphi)$ for all $\theta \in \Theta_1$.

This conjecture does not hold. A counterexample will be obtained by taking $\varphi = \varphi_{\text{sign}}$. In the rest of this section, φ_{sign} and φ_{hyp} will denote the sign test and Fisher's exact test for the testing problem based on the coin-tossing design; φ'_{sign} and φ'_{hyp} will denote the corresponding tests for the alternating design. Remark that $E_\theta(\varphi_{\text{sign}}) = E_\theta(\varphi'_{\text{sign}})$ for all θ on the line-segment joining $(\frac{1}{2}, \frac{1}{2})$ and $(1, 0)$. Moreover φ_{sign} is unbiased size-α (φ'_{sign} is not unbiased!). Hence, if a better test φ' should exist, then this would be unbiased size-α too. But φ'_{hyp} is UMP unbiased size-α. We would get $E_\theta(\varphi'_{\text{hyp}}) \geqq E_\theta(\varphi') \geqq E_\theta(\varphi_{\text{sign}}) = E_\theta(\varphi'_{\text{sign}})$ for all θ on the line-segment joining $(\frac{1}{2}, \frac{1}{2})$ and $(1, 0)$. But we have $E_\theta(\varphi'_{\text{hyp}}) < E_\theta(\varphi'_{\text{sign}})$ for all these θ's (see the beginning of this section). Thus we obtain a contradiction.

In practical situations the experimenter can usually choose between the coin-tossing design and the alternating design before doing anything. No design is better than the other in the above-mentioned theoretical sense. Nevertheless we need a guideline. It follows from our computations (one should try to prove these results), that $E_\theta(\varphi_{\text{sign}}) \geqq E_\theta(\varphi'_{\text{sign}})$ and $E_\theta(\varphi_{\text{hyp}}) \leqq E_\theta(\varphi'_{\text{hyp}})$ holds for all $\theta \in \Theta_1$. The guideline becomes as follows. If n is so small (say $n \leqq 16$) that one would prefer the sign test for both designs, then it is best to use the coin-tossing design. If n is so large (say $n \geqq 40$) that one would prefer Fisher's exact test for both designs, then it is best to use the alternating design.

We expected that the alternating design would constitute a substantial improvement over the coin-tossing design if φ_{hyp} is used. The computations showed that the improvement was not substantial. The expression

$$\sup_{\theta \in \Theta_1} \{E_\theta(\varphi'_{\text{hyp}}) - E_\theta(\varphi_{\text{hyp}})\} \tag{9.1}$$

was about .02 for $n = 30$ and $\alpha \approx .05$; for $n = 10$ and $\alpha \approx .05$ the maximum loss in power (9.1) was about .06.

10. Fisher's exact test is UMP unbiased for the "inexplicable" design. Let D_u denote the class of all unbiased size-α tests, D_s that of all similar size-α tests and D_n that of all tests having Neyman-structure with respect to the statistic $T = (X_1 X_3 \cdots X_{2n-1} Y)$. The result in the title is proved if we can show that (i) $D_u \subset D_s$, (ii) $D_s = D_n$, (iii) φ_{hyp} is UMP (D_n), (iv) $\varphi_{\text{hyp}} \in D_u$.

(i) Introduce the usual topology in the $2^n + 1$ dimensional parameter space $\Theta = \Delta \times \mathscr{P}$ (see Section 8). The power-function $E_\theta(\varphi)$ is a continuous function of θ for each test-function φ. Θ_0 consists of boundary points of Θ_1 only. Hence each unbiased size-α test is similar size-α.

(ii) We must show that T is a complete sufficient statistic for $\{P_\theta;\ \theta \in \Theta_0\}$. If $\theta = (\rho, \rho, P)$ then (8.2) becomes

$$P_\theta(\{x\}) = P[\{(x_1 x_3 \cdots x_{2n-1})\}]\rho^{n-Y(x)}(1 - \rho)^{Y(x)} \tag{10.1}$$

where Y is the rv $S_{00} + S_{10}$ introduced in Section 3. Sufficiency follows from the Factorization Lemma. If $t = (x_1 x_3 \cdots x_{2n-1} y)$ is a possible outcome of T, then

$$P_\theta(T = t) = \binom{n}{y} P[\{(x_1 x_3 \cdots x_{2n-1})\}]\rho^{n-y}(1 - \rho)^y \tag{10.2}$$

because the inverse image $T^{-1}(t)$ contains $\binom{n}{y}$ points x each having the probability (10.1). Completeness of T follows by proving that $E_\theta\{\psi(T)\} = 0$ for all $\theta \in \Theta_0$, implying $\psi = 0$ (ψ is regarded as a function over the range of T). But

$$E_\theta\{\psi(T)\} = \sum_{y=0}^{n} \chi(y)\binom{n}{y}\rho^{n-y}(1 - \rho)^y \tag{10.3}$$

where

$$\chi(y) = \sum_{x_1=0}^{1} \cdots \sum_{x_{2n-1}=0}^{1} \psi(x_1 x_3 \cdots x_{2n-1} y) P[\{(x_1 x_3 \cdots x_{2n-1})\}]\,. \tag{10.4}$$

Now suppose $E_\theta\{\psi(T)\} = 0$ for all $\theta = (\rho, \rho, P) \in \Theta_0$. It follows from (10.3) and the completeness of the family of binomial $B(n, p)$ distributions $(0 < p < 1)$ that $\chi(y) = 0$ for $y = 0, 1, \cdots, n$. But then (10.4) implies that $\psi(x_1 x_3 \cdots x_{2n-1} y) = 0$

because $\chi(y) = 0$ must hold for all $P \in \mathscr{P}$. Hence $\psi(t) = 0$ for all t in the range of T.

(iii) If $\theta = (\rho, \rho, P) \in \Theta_0$, then (10.1) and (10.2) show that all points x in the inverse image $T^{-1}(t)$ have the same conditional probability $P(X = x \mid T = t) = 1/\binom{n}{y}$, where of course $t = (x_1\, x_3 \cdots x_{2n-1}\, y)$. If $\theta = (\theta_0, \theta_1, P)$ is an arbitrary but fixed point in Θ_1, then the above-mentioned conditional probability can be written in the following form

$$c(\theta_0, \theta_1, t)[\theta_1(1 - \theta_0)/\{\theta_0(1 - \theta_1)\}]^{S_{00}(x)} \tag{10.5}$$

as a strictly decreasing function of $S_{00}(x)$ only (under the condition). Hence there exists a UMP (D_n) test, and this conditional level-α test rejects for small outcomes of S_{00} under the condition $T = t$. But this is exactly what φ_{hyp} does. Hence Fisher's exact test φ_{hyp} is UMP (D_n).

(iv) The unbiasedness of φ_{hyp} is trivial because φ_{hyp} is UMP similar size-α.

11. Generalizations. We forced the individual to state whether he prefers the first or the second treatment. It is also possible to allow the individual to declare himself undecided. It is even possible and sometimes attractive to allow the individual to choose one out of, say, *five different categories* like "first treatment much better than second," "first somewhat better," "no preference," "second somewhat better" and "second much better than first." One might try to develop a corresponding theory for the coin-tossing design, the alternating design etc. In our opinion the problems have to be formulated as testing problems where the alternative is restricted by a number of inequalities. We claim that [5] Sections 8 and 10 are of interest. It is also possible that the individual compares (or rather ranks) *more than two treatments* which are administered to him in a certain order. One will expect that a one-sided analogue of Friedman's test arises as a generalization of the sign test. It is also possible that the individual does not provide one score describing his preferences, but that the individual provides *two scores*, one for the first treatment and one for the second. In the rest of this section we shall consider the case when both scores are dichotomous: the individual states for both treatments whether they help or not. We shall see that the corresponding theory is already very intricate and that [5] may be of application.

Our outcome space $\mathscr{X}$ may be described as the space of all 2^{3n} possible sequences $x = (x_1, \cdots, x_{3n})$ of zeros and ones; $x_{3i-2} = 0$ (or 1 respectively) if the ith individual tried the placebo (genuine drug) first; $x_{3i-1} = 0$ (or 1) if the ith individual declares that the first treatment does not help (or that it is a success); $x_{3i} = 0$ (or 1) if the second treatment is a failure (or a success). The probabilistic model will now be based on the conditional probabilities

$$\begin{cases} \theta_{hk} = P(X_{3i-1} = h;\, X_{3i} = k \mid X_{3i-2} = 0) \\ \theta'_{hk} = P(X_{3i-1} = h;\, X_{3i} = k \mid X_{3i-2} = 1) \end{cases} \tag{11.1}$$

where $h, k = 0, 1$ and where of course $\sum\sum \theta_{hk} = \sum\sum \theta'_{hk} = 1$.

It is interesting to try to formulate the null-hypothesis H and the alternative A that the genuine drug is of any help. Of course

$$H: \theta_{hk} = \theta'_{hk} \qquad (h, k = 0, 1), \tag{11.2}$$

but how are we going to formulate the *one-sided* alternative A? In our opinion A has to be defined by means of the following inequalities

$$\begin{cases} \theta_{10} < \theta'_{10}; & \theta_{10} + \theta_{11} < \theta'_{10} + \theta'_{11} \\ \theta'_{01} < \theta_{01}; & \theta'_{01} + \theta'_{11} < \theta_{01} + \theta_{11}. \end{cases} \tag{11.3}$$

This is an alternative, restricted by a number of linear inequalities such that the hypothesis is defined by the corresponding equalities. A theory for such problems was developed in [5].

One might try to find the analogue of the sign test by considering the assumption of *no order effects*:

$$\theta_{hk} = \theta'_{kh} = \pi_{hk} \qquad (h, k = 0, 1). \tag{11.4}$$

By considering the coin-tossing design and assuming (11.4), one arrives at McNemar's test as a UMP unbiased size-α test. We shall not work this out here.

Acknowledgments. The computations were performed on the Telefunken TR-4 of Groningen University, by means of programs written by Mr. L. Th. v. d. Weele. We are indebted to the referees for suggesting improvements in the final form of the manuscript.

REFERENCES

[1] DYNKIN, E. B. (1951). Necessary and sufficient statistics for a family of probability distributions. *Selected Trans. Math. Statist. Prob.* **1** 17–40.

[2] FERGUSON, T. S. (1967). *Mathematical Statistics.* Academic Press.

[3] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[4] National Bureau of Standards (1949). *Tables of the Binomial Probability Distribution.* U.S. Government Printing Office, Washington.

[5] SCHAAFSMA, W. (1966). *Hypothesis Testing Problems with the Alternative Restricted by a Number of Inequalities.* Noordhoff, Groningen.

[6] WITTING, H. (1966). *Mathematische Statistik.* Teubner, Stuttgart.

RIJKSUNIVERSITEIT
MATHEMATISCH INSTITUUT
HOOGBOUW WSN-PADDEPOEL
GRONINGEN
THE NETHERLANDS