COMMUN. STATIST.-THEOR. METH., A5(8), 737-751 (1976)

TESTS OF STATISTICAL HYPOTHESES AND THEIR USE IN STUDIES OF NATURAL PHENOMENA

Jerzy Neyman

Statistical Laboratory
University of California, Berkeley, CA 94720

Key Words & Phrases: immunotherapy of cancer; low-income housing; subsidized drive to unsubsidized homeownership.

ABSTRACT

Contrary to ideas suggested by the title of the conference at which the present paper was presented, the author is not aware of a conceptual difference between a "test of a statistical hypothesis" and a "test of significance" and uses these terms interchangeably. A study of any serious substantive problem involves a sequence of incidents at which one is forced to pause and consider what to do next. In an effort to reduce the frequency of misdirected activities one uses statistical tests. The procedure is illustrated on two examples: (i) Le Cam's (and associates') study of immunotherapy of cancer and (ii) a socio-economic experiment relating to low-income homeownership problems.

1. INTRODUCTION

The title of the present session involves an element that appears mysterious to me. This element is the apparent distinction between tests of statistical hypotheses, on the one hand,

and tests of significance, on the other. If this is not a lapse of someone's pen, then I hope to learn the conceptual distinction.

Particularly with reference to applied statistical work in a variety of domains of Science, my own thoughts of tests of significance, or EQUIVALENTLY of tests of statistical hypotheses, are that they are tools to reduce the frequency of errors. This, of course, makes the theory of testing statistical hypotheses a part of the all inclusive theory of statistical decision functions founded by Abraham Wald and subsequently developed by a great number of colleagues, members of IMS, of ASA, of the Biometric Society, etc.

The achievements of all these scholars differ in many respects. First, peculiarities of domains of empirical studies can generate somewhat different problems of statistical theory. Another kind of differentiation is connected with the unequal richness of the mathematical tool boxes of particular scholars, and here I must confess being envious of Lucien Le Cam's tool box.

The following sections of the present paper are given to two rather diverse examples of current applied research in which tests of statistical hypotheses play an important role. One example is the work of Le Cam conducted jointly with certain associates, specialists in biology and medicine. The ultimate objective of this most interesting study is to develop a novel method of treating cancer, namely through stimulating defense mechanisms of the patient. The other example is the analysis of follow-up data relating to a socio-economic experiment performed some years ago. Jointly with Mark W. Eudey, I am personally involved in this analysis. In both cases only sketches of the background and a couple of details can be reported here.

2. LE CAM'S WORK ON CANCER IMMUNOLOGY

My information on this subject stems from chats with Le Cam and from several talks in our seminar given by Le Cam's associates, Dr. Vera S. Byers, an immunologist, and Dr. Alan S. Levin, an M.D. Briefly and very roughly a small fraction of the study can be summarized as follows.

Apparently, cancer cells removed from a patient can be kept alive in a laboratory. Also, in appropriate conditions, these cells can multiply. It is colonies of such cells that are used in experimentation. One purpose of these experiments is to identify agents capable of killing cancer cells. One such agent is the readily available pure distilled water: within a relatively short time all the cancer cells immersed in distilled water are killed, apparently, irrespective of the kind of cancer. For this reason, the effectiveness of other cancer-killing agents is frequently measured by their cancer cell kill during 3 hours expressed as percentage of that by distilled water. This percentage is labeled "efficiency index".

The suspected (or shall I say, the hoped for) biological cancer killers are certain entities called lymphocytes which are present in our blood. The experiments already performed (Byers, 1975) indicate that lymphocytes taken from individuals of the general population vary substantially in their cancer-killing capacity. This finding is summarized by saying that individuals of the general population vary considerably in their immunity to cancer. (It will be remembered that the immunity in question is what might be called immunity "in vitro", which may or may not parallel immunity "in vivo".)

The other important finding is that the cancer-killing ability of lymphocytes taken from a given person shows a degree of cancer-specificity. For example, individuals whose lymphocytes

are effective in destroying cells of cancer of the bone are relatively rare. Those showing immunity against breast cancer are more frequent, etc. Thus, it is appropriate to speak of cancer type-specific immunity of particular individuals.

The establishment of the above facts brought out the question about the origin of the cancer type-specific immunity. What about the possibility that such immunity originates from mild exposures to particular types of cancer? The "mild exposure" contemplated consists in living in a household containing a patient suffering from a specified type of cancer. It is labeled "household contact's exposure".

In order to study the possible effects of household contacts, specifically with regard to cancer of the breast, two samples of individuals were examined, and I am indebted to Le Cam for showing me the data in advance of their publication. One of the samples is composed of 34 "control" persons who had no known contacts with women suffering from the cancer of the breast. The other sample contains 33 individuals, labeled the "contact individuals", who, for at least two years, had household contacts with a breast cancer patient.

Each of the 67 persons studied yielded a supply of lymphocytes and these were tested for their breast cancer cell killing ability as measured by the efficiency index. This I denote by X for controls and by Y for contact individuals. The hypothesis to be tested, say H, is clearly indicated by the purpose of the study: the population distribution of Y is identical with that of X. The alternative hypothesis $\bar{H}$ against which the test to be used had to be particularly powerful is also clear. The purpose of the study was to find whether, by and large, the household contacts tend to increase the immunity against cancer of the breast. Evidently, in the affirmative case, the random variable Y would be "stochastically larger" than X. Accordingly, Le Cam's choice was the Mann-Whitney test rather than, say the χ^2 or the Kolmogorov-Smirnov test.

The application of the chosen test left little doubt that the lymphocytes from household contact persons are, by and large, better breast cancer cell killers than those from the controls. The situation is illustrated in Figure 1, indicating the empirical cumulative distributions of X and separately of Y. It will be seen that, as it happens, almost any statistical test worthy of the name would find the two distributions significantly different.

Incidentally, the sample of household contact individuals contained persons genetically related to patients (children, sibs, etc.), and also those without such connections (husbands, adopted children, maids, etc.). One of the additional questions studied in the investigation was whether genetic connections made a difference in the performance of the lymphocytes. No evidence was found. Neither was there any evidence of a difference between lymphocytes of men and women.

I find the study most interesting and am looking forward to seeing the full account of its findings.

3. FOLLOW-UP STUDY OF LIHD-2, A SOCIO-ECONOMIC EXPERIMENT

In the 1960's the Federal Department of Housing and Urban Development (HUD) funded several low-income housing "demonstration" projects undertaken by local groups in the United States. In one particular instance, an effort was made to turn the "demonstration" into a randomized experiment. This was done by a non-profit corporation, the San Francisco Development Fund (SFDF). The purpose of the experiment, labeled LIHD-2, was to find out whether in the general category of low-income families a "select" group could be found for which counseling and modest limited-duration subsidies would lead to improvement in housing and, eventually, to stable home ownership. The operational period of the experiment was December 1966 through November 1969. It involved 104 families, 52 experimentals and 52 controls. The intention was to form 52 ethnically matching pairs of families,

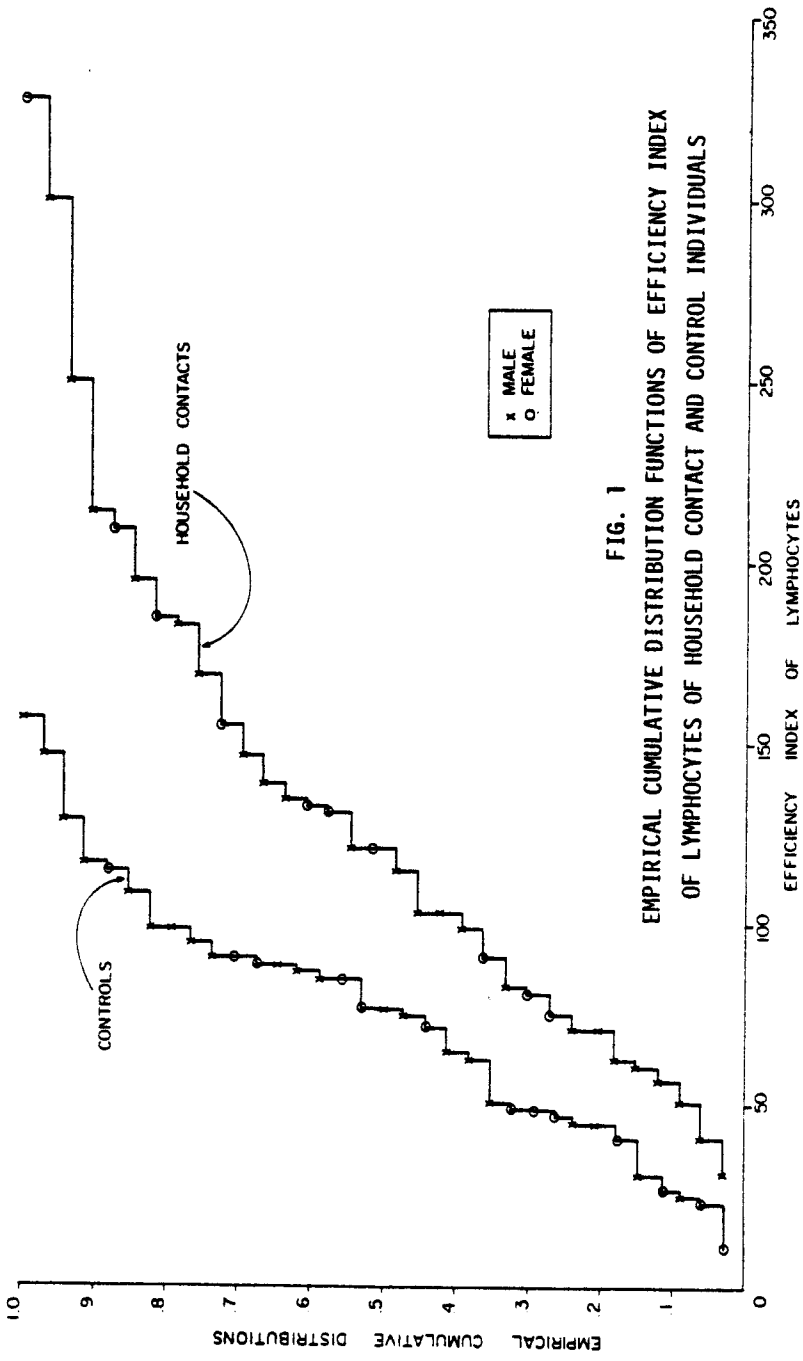


FIG. 1
EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTIONS OF EFFICIENCY INDEX OF LYMPHOCYTES OF HOUSEHOLD CONTACT AND CONTROL INDIVIDUALS

one of which, selected at random, would be assigned to the experimental category and the other to the controls. As will be seen below, this most laudable goal proved difficult to achieve.

In order to be included into the experiment, the recruited families had to pass a test which consisted in earning a specified score on each of certain six qualifying criteria devised in conferences with sociologists. These criteria included the so-called "potential", expected to measure the capability of the family to rise above preexisting cultural and economic levels, to increase its earnings, etc.

After being assigned to the experimental group, each qualified family was moved to a reasonable rented apartment and was supplied with subsidies averaging about $50.00 per month and continuing for no more than two years. This two-year period served the families to become accustomed to better housing (with enough bedrooms, etc.), to learn the technicalities of home acquisition, to find ways of increasing their incomes (wives proved very helpful), to budget their earnings and expenses and to shop for a desirable home to buy. The staff of SFDF tried to help by counseling. All this applied to the experimental families. As to the control families, all the benefits they got from being included in the experiment were limited to the hypothetical stimulus from being "qualified" and to a $10.00 fee for an interview at the end of the experiment.

When the experiment was concluded, in November 1969, it was found, not very surprisingly, that the experimental families advanced to homeownership much faster than the controls (Eudey, 1970). However, there was the question about the durability of the successes. Therefore, three and a half years after the conclusion of the experiment, in April-May, 1973, the SFDF performed a follow-up study, which is the subject of the present communication.

I must begin with the coverage. The follow-up study could have been included in the original plan of the experiment, and

then arrangements might have been made to keep all the 104 families more or less in sight. As it happened, the organization of the follow-up began later, and the field work was limited to two months. The result was that out of the 104 families only 97 could be located, 48 experimentals and 49 controls. While the loss of seven families is regrettable, it must be clear that their availability could not alter very much the general picture provided by those located and interviewed. The difficulty of the follow-up analysis is elsewhere.

[Remark: This is not the case with the recent survey (Urban Management Consultants of San Francisco, 1975) of two low-income housing programs, one of which was administered by SFDF. This year-long investigation was intended to provide the Federal Government with information on the effectiveness of two different policies. The samples of families covered by the survey are not negligible, 144 in one case and 473 in the other. However, important comparisons are made with unexpectedly large proportions of missing data. For example, the foreclosure rates are compared on data with 22% missing records for one program and with 28% missing for the other. Similarly, the comparison of frequencies of "late fees" among the same groups of families is made with omissions of 34% and 46% of families concerned, respectively.]

As mentioned in the description of LIHD-2 (Eudey, 1970), the strictly random partitioning of all the qualified families into experimental and control groups proved difficult. At the time, the bias favoring the experimental category appeared small. However, when after the follow-up all the data were reexamined, it became clear that the bias is quite important. Specifically, the experimental category contained substantially more families with relatively high scores on "potential" than the control category. The same applies to the score on "education". Thus, the idea of analysis through a comparison between the experimental and control families was abandoned. The whole analysis is bulky and we

hope to publish it elsewhere. Here only one detail must suffice. It is concerned with the selection of a statistical test to validate one of the basic ideas of the experiment, namely that, in conditions of the experimental families, high scores on "potential" and on education do indicate good prospects in a move towards unsubsidized homeownership.

Our reasoning, Mark Eudey's and mine, is as follows. Consider the group of 48 experimental families located at the follow-up and divide it into two equal subgroups, those with potential score below the median (say "low P" or LP) and those above (say HP). Next, we visualize two populations of which the LP and the HP families can be considered as random samples. Let $\Pi(L)$ and $\Pi(H)$ denote those populations. The question about the relevance of "potential" is now reduced to the question whether the probability distribution of success in the move to homeownership within the population $\Pi(H)$ coincides with that within $\Pi(L)$. After realizing this, we have to face the problem of defining a measure of success. Here one point is obvious. This is that, if at the time of the follow-up a family does not own its home, then its success, say S_0, is zero. But what about those families that in the spring of 1973 were found to be homeowners? Are they all to be treated equally? We have several thoughts on this matter, one of which is as follows.

The follow-up data contains information not only on whether, at the time, a family lived in its own home but also on the date at which this home was purchased. It appears to us that this date is somehow relevant to the degree of success achieved by the family. It will be remembered that each experimental family had a period of up to two years of subsidies during which it could acquire quite a bit of technical information and of education in matters relating to its income, to budgeting, etc. One way of thinking that appeals to us is that the purchase of a home during the period of subsidies, even if this home is still owned in 1973, may be a stroke of good luck combined with somewhat

excessive enthusiasm. We assign to it what may be called "moderate success" and denote it by S_1. This is contrasted with the success of the family found living in its home purchased after the conclusion of the period of subsidies. The performance of such a family appears impressive to us because it demonstrates its ability to increase its earnings and to budget its expenses sufficiently to be able to produce the downpayment entirely on its own. We call its success "high" and denote it by S_2.

The other way of thinking leads to a reverse use of labels "modest" and "high" success. Here the basic consideration is that, if a family owning a house in 1973 bought it more than three years before, then this family demonstrated a substantial degree of homeownership stability, which is rather an important characteristic. Here, then, the early purchase of the home still owned in April-May of 1973 appears to deserve the label of "high" performance, S_2.

With S_0 unambiguously defined and with two measures of success S_1 and S_2 open to some subjective discussion, let us now face the problem of selecting an appropriate test of the hypothesis, say H, that the probabilities of the three degrees of success in populations $\Pi(L)$ and $\Pi(H)$ are the same. The basic idea was proposed some years ago by Mrs. Dorothy Marshak, then our graduate student. It was conceived in connection with our study of the existence or non-existence of a physical phenomenon of "memory boost". The deduction of the optimal $C(\alpha)$ test is due to F.N. David (1972). The general scheme is as follows.

Denote by θ_0, θ_1 and $\theta_2 = 1 - \theta_0 - \theta_1$ the unknown probabilities of successes S_0, S_1 and S_2 corresponding to $\Pi(L)$ families, with such definition of S_1 and S_2 as fits the reader's intuition best. If the score on "potential" (or on "education") is really meaningful, then (a) the probability of S_0 in $\Pi(H)$ ought to be smaller than that in $\Pi(L)$; and (b) the probabilities of S_2 should be in reverse relation. Accordingly, Mrs. Marshak suggested a set-up as in Table I, where ξ denotes a non-negative number less than unity.

TABLE I

Probabilities of Successes in $\Pi(L)$ and $\Pi(H)$

Success	S_0	S_1	S_2
$\Pi(L)$	θ_0	θ_1	θ_2
$\Pi(H)$	$\theta_0(1-\xi)$	$\theta_1 + \theta_0\xi - \theta_1\xi$	$\theta_2 + \theta_1\xi$

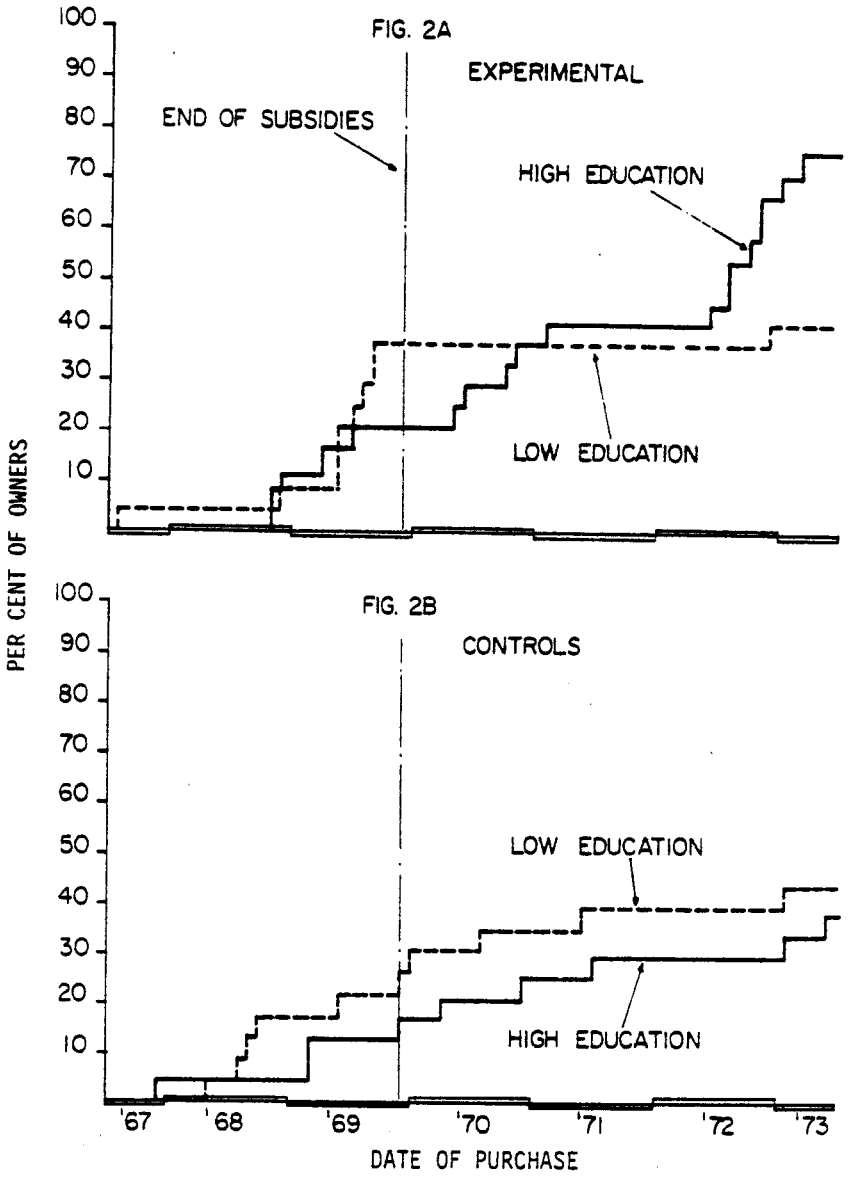
With this set-up, the null hypothesis H to be tested reduces to the assertion that $\xi = 0$, the alternative being $\xi > 0$. Table II gives the numbers of experimental families, located at the follow-up, classified according to their "high" and "low" scores on "potential" and on "education". Here, the first of the two definitions of S_1 and S_2 is used.

The application of the test left little room for doubt that the score on potential and also that on education are very relevant to the prospects that a "qualified" low-income family put in conditions of the "experimentals" will be successful in its move towards non-subsidized home acquisition. Figures 2A and 2B, one referring to experimental families and the other to controls,

TABLE II

Performance of Families with "high" and "low" Scores On "potential" and "education"

Measure of success	S_0 = not owner at follow-up	S_1 = owner, bought during subsidy period	S_2 = owner, bought after subsidy period
High P	5	10	9
Low P	15	4	5
High E	6	5	13
Low E	14	9	1
Totals	20	14	14



FIGS. 2A, B

CUMULATIVE CHRONOLOGIES OF AQUISITION OF HOUSES OWNED AT FOLLOW-UP

both classified on their scores on education, provide an intuitive feeling of what must have been going on. I am particularly impressed by the finding that, with a single exception, the experimental families with lower than the median score on education, bought their homes before the expiration of the period of subsidies, while those with "high" education continued their buying spree up to the end of the period of observation, in 1973.

Figure 2B, relating to control families, presents a picture entirely different from that in Figure 2A. While the comparison of the two figures is interesting, it will be remembered that the striking difference is ascribable not only to the treatment received by the experimentals but also to the fact that, because of the bias in assignment, the group of controls is somehow weaker than the experimentals.

A POSTSCRIPT: REMARKS SUGGESTED BY THE DISCUSSION AT THE MEETING

(i) In the post Abraham Wald terminology, the problem of testing a statistical hypothesis H is a "two-decision problem": (a) either reject H or (b) do not reject H. The phrase "do not reject H" is longish and cumbersome and, therefore, there are several alternatives in common use. One is "accept" $\bar{H}$, which is the negation of H or its "alternative". My own preferred substitute for "do not reject H" is "no evidence against H is found".

The two-decision problem of testing H should be distinguished from a "three-decision problem". Here, the three alternative actions are: (a) accept H, (b) accept $\bar{H}$, and (c) remain in doubt.

The theoretical statistical problems relating to the above situations consist in developing criteria which tend to minimize the frequencies of errors, particularly those subjectively considered as especially important to avoid.

(ii) Problems of estimation, whether point estimation or by interval, are multi-decision problems (except when the quantity to be estimated can have one out of two specified possible values).

(iii) It seems to me that the distinction between "tests of significance" and "tests of statistical hypotheses" mentioned in the title of the Session is not really one applicable to some two conceptual entities but to subjective attitudes of the practicing statistician. Thus, the same statistic, say the "Student"-Fisher t-statistic, may be used in two different capacities determined by the attitude of the user. One seems to be the capacity of a "test of significance" and the other that of a "test of a statistical hypothesis". In my own empirical work on a "substantive" problem such distinctions do not appear necessary. My use of the t-statistic is limited to certain familiar conditions in which its frequency properties have been proved to be, in a sense, optimal.

(iv) A similar remark applies to the use of the words "decision" or "conclusion". It seems to me that at our discussion these particular words were used to designate only something like a final outcome of complicated analysis involving several tests of different hypotheses. In my own way of speaking, I do not hesitate to use the words "decision" or "conclusion" every time they come handy. For example, in the analysis of the follow-up data for the LIHD-2 experiment, Mark Eudey and I started by considering the importance of bias in forming the experimental and control groups of families. As a result of the tests we applied, we decided to act on the assumption (or concluded) that the two groups are not random samples from the same population. Acting on this assumption (or having reached this conclusion), we sought for ways to analyze that data other than by comparing the experimental and the control groups.

The procedures I described involved tests of two new hypotheses, namely that "high" or "low" scores on "potential" and on "education" do not affect the chances of success in the drive to home ownership. The analyses we performed led us to "conclude" or "decide" that the hypotheses tested could be rejected without excessive risk of error. In other words, after considering the

probability of error (that is, after considering how frequently we would be in error if in conditions of our data we rejected the hypotheses tested), we decided to act on the assumption that "high" scores on "potential" and on "education" are indicative of better chances of success in the drive to home ownership.

I need not emphasize that this substantive conclusion applies to a particular locality, the San Francisco Bay Area, and to the particular epoch, late 1960's and early 1970's.

ACKNOWLEDGMENT

This paper was prepared using the facilities provided by Grant GM 10525 of the National Institutes of Health.

BIBLIOGRAPHY

Byers, V.S., Levin, A.S., Hackett, A.J. & Fudenberg, H.H. (1975). Tumor-specific cell-mediated immunity in household contacts of cancer patients. J. Clinical Investigations 55, 500-13.

David, F.N. (1972). Applications of Neyman's C(α) technique. Proc. 6th Berkeley Symp. Math. Statist. Prob. IV. Berkeley: University of California Press, 223-30.

Eudey, E. (1970). A Move to Homeownership. Publication of the San Francisco Development Fund, Inc., 57 Port Street, San Francisco, CA 94104.

Urban Management Consultants of San Francisco, Inc. (June 1975). Evaluation of the San Francisco Development Fund's Buyer's Agent Program. Report on Contract No. H-220R.

Received January 1976; retyped February 1976.