APPENDIX.

PLANNING AN EXPERIMENT WITH CLOUD SEEDING

JERZY NEYMAN and ELIZABETH L. SCOTT
UNIVERSITY OF CALIFORNIA, BERKELEY

A1. Introduction

The general principles of the theory of experimentation were established by R. A. Fisher in his memorable book *The Design of Experiments* [1], and then developed by Fisher's innumerable followers and coworkers of whom we shall particularly mention Frank Yates [2]. Primarily, the designs studied by these scholars refer to experimentation in biology and, more particularly, in agriculture. Naturally, each domain of experimentation presents certain particularities and, while the general principles of experimental design remain the same for all domains, each particular domain imposes its own limitations and requires special designs. In particular, experimentation with clouds or storms involves specific difficulties not encountered, for example, in the experiment with the Lady Tasting Tea, the famous problem used by Fisher to illustrate his ideas.

Each of the experiments reported in these *Proceedings* involved a substantial amount of planning and the experience gained will be most useful in designing future experiments. The purpose of the present appendix is to contribute to this goal by reviewing the problem as a whole and by focusing attention on several subproblems which appear to us of particular importance. This should be done with reference to as many already completed experiments as possible. Unfortunately, our familiarity with quite a few of these experiments is of a very recent date and some of the very important problems raised by them, including the problem of possible after effects of seeding noticed in Australia, must be left out of consideration.

Numerical illustrations given below are based, predominantly, on data collected by project SCUD [3]. Here, the achievements of the planners, a meteorological group headed by Dr. Jerome Spar cooperating with a statistical group headed by Dr. John W. Tukey, seem to have received less attention than they deserve.

Prepared with the partial support of the Office of Naval Research (Contract No. N00014-66-C00036-G01; NR 307-303X) and the U.S. Army Research Office (Durham) DA-31-124-ARO-D-548.

In general, planning an experiment has two aspects: the substantive, in our case meteorological, and the statistical.

A2. Meteorological aspects of planning an experiment with cloud seeding

The meteorological aspects of planning an experiment with cloud seeding depend upon past experience, upon what the experimenter is prepared to adopt as a working hypothesis and upon the questions that one wishes to have answered by the experiment.

Most of the experiments reported in this volume appear to have been designed with the idea that the seeding of clouds by particular methods favored by particular experimenters, the seeding in some specified conditions visualized by these experimenters as particularly favorable, will produce a specified desired effect: an increase in precipitation in some cases, suppression of hail or lightning in some others, and so forth. The purpose of such an experiment was, invariably, to prove that the ideas of the experimenters concerned are correct.

This attitude, "to prove" occasionally contrasted with the statements that the purpose of the experiment is to explore, is manifested in the tendency to define in advance the presumed conditions of "seedability" and to select for experimentation only such units (storms, and so on) that conform with the given definition.

The following quotations illustrating the attitude "to prove" must suffice.

(i) ". . . designs of future experiments which would . . . give definite evidence as to the rainfall increases which may be achieved."

(ii) "Experiments were performed only on cumulus clouds which complied to a fixed specification; they had to be supercooled, reasonably isolated, deep, of long duration, and without excessive sheer, and not within 30 km of any other cloud which is raining or glaciated."

(iii) "Cumulus clouds with tops warmer than −10°C will not be seeded."

In addition, the difference between the two randomized experiments in Arizona might be mentioned. Here, after four years of experimentation, it was decided to continue with a stricter selection of experimental units.

The record of experimentation assembled in the present *Proceedings* seems to indicate that guesses as to the conditions in which cloud seeding by the different methods would produce this or that hoped for effect were frequently unsuccessful. In most American experiments it was hoped to demonstrate that the seeding will increase precipitation. Contrary to this, there is evidence that, if anything, the precipitation was decreased. The purpose of the Swiss experiment Grossversuch III was to demonstrate that hail may be suppressed by seeding. Contrary to this, Paul Schmid produces evidence [4] that the frequency of days with hail was increased by seeding. At the same time, our own analysis of the Swiss data brought out the fact that, at least in some sections of the target in Switzerland, the overall effect of seeding on precipitation was a significant and sizable increase.

In these circumstances, the realistic working hypothesis to adopt in planning a future experiment is that cloud seeding can affect both hail and precipitation in one way or another, probably depending upon the method of seeding and upon local orographic conditions, and to stop at that. With this working hypothesis, the object of an experiment would be precisely to determine the conditions under which cloud seeding has this or that effect. Here, the loosely used term "conditions" refers to three different categories of factors: method and rate of seeding, locality of the experiment (determining the orographic conditions), and the synoptic situation.

Because of the particular purposes of experimentation, the policy behind future experiments should be contrary to that reflected in some of the above quotations. Rather than restrict the experimenting to occasions satisfying some preconceived conditions, perhaps to those when the "precipitable water" exceeds a certain limit, or when cloud tops are colder than a given limit, and so forth, the future experiments should include cases with conditions varying within the broadest possible limits, perhaps all cases where there is any appreciable chance of natural rain.

What should be the definition of these "broadest possible" conditions is, of course, a strictly meteorological problem. It is mentioned here as a matter of general principle, analogous to the familiar statistical result that, in order to be able to estimate the regression of Y on X, the nature of which is not known *a priori*, one should observe Y with reference to X varying within broadest possible limits. In retrospect, it is not an accident that the only unambiguous indication that seeding with silver iodide can increase rain comes from Grossversuch III. The reason is that in this particular experiment the forecasters were concerned with predicting thunderstorms and nothing else. Also, it appears that the several forecasters concerned had somewhat different tendencies in their work. As a result, the seven year long experiment accumulated a large number of 292 experimental days, about 42 per year, and these days involved a variety of weather patterns, apparently conditioning different effects of seeding.

To be informative, the suggested inclusive program of seeding must be accompanied by an extensive program of collateral observations on all those factors that may conceivably be related to the conditions of seedability. James Hughes' idea of winds aloft (see figure 1 in the main body of this paper), Jerome Spar's invention [3] of his predictor variables and considerations of temperature and humidity of the upper air indicate the desirability of radiosondes, presumably sent out locally and more frequently than is done in routine observations. Also, Henderson's report [5] on unpredictability of silver iodide plumes indicates that similar observations might be very useful in future experiments. Undoubtedly, many other observations will be indicated by knowledgeable meteorologists.

Given a comprehensive experimental program, involving a broad variety of meteorological conditions, and given an equally broad program of collateral observations, there will be a meteorological problem to use the latter in order to split the totality of meteorological conditions into a reasonable number of categories, hopefully each category determining a different effect of seeding. What these categories should be is a matter for cloud physicists to determine and here the experience with already completed experiments is likely to be very useful. Just as a matter of illustration, the experience with Grossversuch III indicates the relevance of winds aloft, of thus far vaguely defined "incipient" and "dissipating" storms and of a particular kind of storm situation described as "Barrage" ([4], table III). These identifiable categories of weather patterns would be included in a comprehensive evaluation designed to test the tentative hypotheses underlying the definitions of the categories.

Compared with the old style experiments, characterized by the attitude "to prove," the proposed experiment would be substantially richer. In fact, it would include *all* the experimental units (experimental days, or storms) as would an old style experiment, but, in addition, it would include other units which in an old style experiment would not have been included because of the lack of compliance with the preset fixed specifications. The subsequent evaluation would not be concerned with a single summary overall effect of seeding, but with *effects* of seeding (we emphasize the plural) in the different sets of conditions included in the experiment. This, then, will implement the attitude "to explore" contrasted with that "to prove."

While advocating comprehensiveness with regard to weather patterns, we would like also to advocate similar comprehensiveness with regard to methods of seeding and of orographic conditions. With regard to the latter, the situation is simple. In order to investigate the effectiveness of seeding in several different sets of orographic conditions, it is unavoidable to organize several different experiments. With regard to testing several different methods of seeding in some fixed orographic conditions, the situation is more complicated. One is tempted to organize a single experiment with several methods of seeding, as is successfully done in agriculture and elsewhere. However, such simultaneous study of several treatments involves a risk which should be taken into account. Suppose, for example, that in a single experiment it is desired to study the effectiveness of some two methods of seeding, A and B. Unavoidably, this would involve splitting the totality of experimental units into at least three, and possibly four, categories: controls, without any seeding, units with seeding by method A, those with method B and, possibly, those with both methods A and B being used simultaneously. This is a method, labeled factorial design, that appears very efficient in biology and that may prove efficient in cloud seeding experimentation. The limitation is the possibility of interaction between the treatments A and B. That is, if the effectiveness of B in the presence of A is not (approximately) the same as in the absence of A, then, in order to have a reasonable chance of detecting the effectiveness of either A or B the duration of the experiment as originally planned may not be sufficient. Naturally, dangers of this kind exist in agricultural experimentation and elsewhere. However, in weather modification experiments these dangers are more important because, as the current experience indicates, an informative experiment with cloud seeding must take several years where as an ordinary agricultural trial takes just one season. Also, the costs involved are not comparable. Here, then, there is a difference in the desirability of comprehensiveness. The comprehensiveness with regard to weather conditions would not increase the duration of the experiment, though, of course, it would increase the amount of labor expended. Contrary to this, an attempt at comprehensiveness in the variety of treatments may increase the duration of the experiment. On the other hand, if one adopts a strong limitation on the study of methodology of seeding, one is in obvious danger of putting all the eggs in a single basket and of having this basket smashed. Here, then, every decision involves a risk and the only safe prescription that one can formulate is that it be a "calculated risk."

With the motivation of learning how to increase rain by seeding, it is natural to include in future experimentation those methods for which unambiguous evidence exists that they did increase precipitation. In Grossversuch III the seeding was done with silver iodide from ground based generators.

A3. Statistical aspects of planning an experiment

The basic concept in any experiment involving substantial and unpredictable variation is that of an experimental unit. In biology the experimental unit may be a rat. In rain modification experiments this may be a conventionally defined "storm" extending perhaps from a few hours to several days, or some period of fixed duration, perhaps 12 or 24 hours, for which some rain is forecast, and so forth. Whatever the actual definition of the experimental unit, frequently it will be convenient to speak in terms of storms.

The basic question which an experiment is expected to answer may be formulated in several ways, of which the simplest is as follows.

We visualize a certain number N of experimental units, say $u_1, u_2, \cdots, u_N$, which may be subject to experimentation. These may be storms passing over the given locality during the next five years. Alternatively, these may be storms satisfying certain conditions of seedability, and so forth. With each forthcoming storm, say u_n, we associate a number, say v_n, representing the average precipitation to be recorded in a certain number of raingages in the target area, the precipitation delivered by the storm u_n without seeding. Next, we visualize another number, say w_n, representing the similarly calculated average precipitation from the storm u_n with seeding. Thus, we consider two sequences of numbers

$$v_1, v_2, \cdots, v_n, \cdots v_N \tag{A3.1}$$

$$w_1, w_2, \cdots, w_n, \cdots w_N \tag{A3.2}$$

and their averages $\bar{v}$ and $\bar{w}$, respectively. The simplest (but the narrowest) question that we would like to have answered is whether $\bar{w}$ equals $\bar{v}$ or not. If $\bar{w} = \bar{v}$, then we would say that, in the conditions of the experiment, the average effect of seeding is nil. If $\bar{w} > \bar{v}$ we would say that, on the average, the seeding increased the precipitation, and so forth.

Unfortunately, this fundamental question, as to whether $\bar{w} = \bar{v}$, cannot be answered. It is easy to learn the values of all the numbers v_n of the sequence (A3.1): just refrain from seeding and service the raingages. Then $\bar{v}$ could be calculated without error. It is equally easy to learn all and every number w_n of sequence (A3.2). However, the important thing is that the numbers v_n and w_n cannot both be known for the same nth storm and that, therefore, it is impossible to learn the exact values of both $\bar{v}$ and $\bar{w}$. All that is possible is to seed some storms, thus learning some of the numbers w_n of sequence (A3.2), and let go without seeding the remaining storms of the period, thereby learning some of the numbers v_m of sequence (A3.1), always with $m \neq n$. Then the samples of the v and the sample of w could be used in order to form judgments about the unknowable $\bar{v}$ and $\bar{w}$. It is here that mathematical statistics comes in, but even so there are important limitations.

A scientist would like to know the values of $\bar{v}$ and $\bar{w}$ referring to his own experiment just performed. Contrary to this, with reference to any particular experiment, mathematical statistics, or indeed any other discipline, is powerless to determine whether $\bar{v} = \bar{w}$. On the other hand, methods of mathematical statistics are helpful with reference to long series of experiments. These may be experiments all of the same kind, say cloud seeding experiments, or all different, provided they satisfy the important condition of being randomized. This means that the observational units must be assigned to treatments (for example, the storms to be seeded or not) using some mechanism that insures a preassigned probability for each observational unit to be given this or that treatment and preassigned dependence properties of such assignments of treatments. In particular, each forthcoming storm may be given the same probability, say one half, of being seeded and the seeding of one storm may be made independent of whether the preceding storm was seeded or not. This is what is called unrestricted randomization. Alternatively (and this is preferable), the succeeding storms may be randomized in pairs, and so forth.

Given the randomization, statistical theory provides methods of dealing with experimental data which refer to two related, but different problems. One is the problem of testing hypotheses [6], providing answers to such questions as: should one assume that $\bar{w} = \bar{v}$? The other problem is that of statistical estimation [7], [8] attempting to answer the questions such as the following: (i) if one has to assume some value ξ for the difference $\bar{w} - \bar{v}$, what is the "best" assumption to make? Also (ii) since the true value of $\xi = \bar{w} - \bar{v}$ is unknowable, can one at least use the observations, say X, in order to assign to ξ two bounds, say $\xi_1(x)$ and $\xi_2(x)$ such that the probability that these two bounds will cover the true ξ, so that $\xi_1(x) \leqq \xi \leqq \xi_2(x)$ has a calculable high value? Also, if more than one pair of bounds $[\xi_1(x), \xi_2(x)]$ are available, can one find the one that is the narrowest? This is the problem of confidence limits, [8]. (See also [9].)

The following are the important properties of methods of testing hypotheses.

(i) In a long series of experiments in which the tested treatment has no effect, so that $\bar{w} = \bar{v}$, the frequency of cases where the method leads to the fallacious judgment that $\bar{w} \neq \bar{v}$ has a preassigned value α, such as $\alpha = 0.1, 0.05, 0.01$, and so forth.

(ii) In a long series of experiments in which the effect of the tested treatment exists, and where $\bar{w} - \bar{v} = \xi \neq 0$ the frequency of cases, say $\beta(\xi, \alpha)$ in which the method will bring out the judgment that $\bar{w} \neq \bar{v}$ is calculable.

As is well known to most experimenters with weather modification, the preassigned α is called the level of significance. The probability $\beta(\xi, \alpha)$ called the power of the test, is somewhat less familiar. Its study was introduced in [6].

When the problem of testing a statistical hypothesis is solved, so that an appropriate test is available, the level of significance α can be chosen in advance and the actual frequency of false assertions $\bar{w} \neq \bar{v}$ will be, at least approximately, equal to α, whether the number of observations, say N, is relatively small or large, and irrespective of the various details of the experiment. Ordinarily, this is not true of the power. When α is fixed and the use of the test is adjusted to insure the frequency of erroneous judgments that $\bar{w} \neq \bar{v}$ be approximately α, the value of $\beta(\xi, \alpha)$ depends very much both on the number N of observations and on the various details of the experiment. Also, and this must be intuitively clear, for fixed ξ and fixed conditions of the experiment the smaller the adopted level of significance α (one would naturally want it to be small!), the smaller the power $\beta(\xi, \alpha)$. On many occasions this circumstance places the experimenter, so to speak, between a Scylla of low chance of detecting the effects that his experiment is supposed to detect, and a Charybdis of a high risk of asserting the existence of the effects of treatment when, in fact, these effects are absent. In an effort at caution, the experimenter may wish to insist on a rather strict level of significance, say $\alpha = 0.01$. Then it may happen that, with the fixed conditions of the experiment, including the number of observations, the chance of detecting the increase in precipitation if it is as large as, say, 40 per cent, is very low, perhaps only $\beta = 0.2$. By readjusting the statistical test the value of β may be increased say to $\beta = 0.8$. However, invariably, this would lead to an increase in α, frequently to an inordinate increase.

It follows that the question about the attainable values of α and β must be asked at the time when the experiment is planned. This would involve a statistical study of climatological conditions in and about the prospective target. The study, somewhat analogous to uniformity trials familiar in agricultural experimentation, would run on the general lines of the investigation by Changnon and Huff [10] published in this volume. Broadly speaking, the purpose of such studies is to provide the information necessary for the calculation of simultaneous values of α and β, depending upon the various possibilities in defining the experimental unit and the target. The basic question here is about the distribution of the unseeded rainfall in the target, and perhaps also in some control areas, for variously defined experimental units and for several conceivable modifications in the target, all this in conjunction with some predictor variables. The following illustrations must suffice.

(i) Huff and Changnon provide the rainfall distribution per "storm" defined in a particular manner. Here, then, the storm is a contemplated experimental unit. The published data are then used to investigate the performance of a particular test of effectiveness of seeding. While this is very valuable information, it could be usefully supplemented by similar results relating to units of observation defined differently, perhaps as 24 hour periods. As is well known, one of the main difficulties of experimentation with rain stimulation is the notorious variability in natural rainfall. As defined by Changnon and Huff, the concept of a storm is quite narrow and, frequently, during a period of 24 hours there will be not just one, but several "storms." Intuition suggests that if one adopts as the unit of observation the 24 hour period, rather than a "storm," the variability of precipitation amounts might be decreased. However, this intuitive idea must be verified empirically.

REMARK. It is hoped that the Changnon and Huff data will be used to answer the above and several other similar questions.

(ii) The variability of precipitation per observational unit must depend upon the target chosen and also upon the number of raingages in the target. Among other things, the second of the two experiments in Arizona differs from the first by size of the target chosen and by the density of gages: in the second experiment the target was smaller than in the first and the density of gages was higher. When the experimental results are published, it will be very interesting to examine whether this contraction of the target increased the power of the test.

(iii) The experiments already performed differ in the following important characteristic of the target chosen: some targets are fixed areas, large or small. For example, two targets in SCUD were fixed. The same was true in Arizona and in Switzerland. In other cases the "target" was adjusted separately for each experimental unit, according to the prevailing wind direction, and so forth. This was the case in the Washington-Oregon and in the White Top experiments. Convincing *a priori* arguments might be adduced in favor of either arrangement. However, this is a question of empirical conditions and a reliable answer may be obtained only through an empirical study. If the general area of an experiment is uniform with regard to the rain pattern, a target adjustable to wind direction and velocity is likely to be preferable to a fixed target. On the other hand, if the general area of an experiment includes a variety of orographic conditions with widely different normal precipitation amounts, a reasonably large fixed target may be preferable.

(iv) The already completed experiments differ considerably in the use of predictor variables. In Arizona and in Switzerland no predictor variables were used. In many other experiments, in the United States and abroad, it was more or less customary to use rain in some control areas as predictor variables. In Project SCUD three nonprecipitation predictor variables were used.

The effectiveness of predictor variables in contributing to the precision of an experiment is illustrated below in some detail. Here we wish to point out the desirability of including the search for predictor variables into the planning phase of the experiment. Studies of this kind, as well as studies indicated above under (i), (ii), and (iii), may be hampered by the low density of raingages in an area contemplated for the future experiments. However, a few raingages are likely to be found operating routinely over several years in the past and an effort to use these in planning an important experiment is fully justified.

The importance of reliable empirical information about the distribution of target precipitation from one experimental unit to the next is connected with the existence of a variety of statistical tests, all possible to adjust to the same arbitrarily chosen level of significance. One particular category of such tests, labeled nonparametric, maintains the level of significance irrespective of the underlying distribution of the observable variables. However, these tests differ considerably in their power which depends very much on the distribution of precipitation and on the character of the possible effect of seeding. The situation is illustrated in figure 6 in the main body of this paper, p. 320. Depending upon the test used, always with the same intended level of significance $\alpha = 0.1$, the probability of detecting the effect of seeding when this amounts to a 40 per cent increase varies from about 0.22 to about 0.45. In other conditions, the effectiveness of the tests considered might be the reverse. Thus, when planning the experiment it is essential to have some idea of the distributions likely to be encountered so as to be able to select the statistical test with the greatest possible power and, finally, in order to estimate the number of experimental units, or the duration of the experiment, which would insure a reasonable chance of detecting the effects that one wishes to detect.

Before concluding this section, we must return to randomization as a prerequisite to planning a proper experiment. As is well known, ordinarily, commercial cloud seeding experiments are not randomized and yet their results are being subjected to statistical evaluations and reevaluations. The question may be asked why should randomization be considered so necessary for what we call "experiments."

For a detailed discussion of the question we cannot do better than refer the reader to the book by Fisher [1] already quoted. Here it must suffice that, without randomization, any apparent effects of the experiment, no matter how "statistically significant," cannot be unambiguously attributed to seeding. Reliance on nonrandomized operations places the scientist in the position of the Court of King Arthur so easily misled by the indomitable Connecticut Yankee [11].

The very possibility of being misled should be sufficient to concentrate the attention on randomized experiments where, with proper care, misattribution of the noted effects is impossible. However, it is no doubt interesting to put one's finger on some actual source of bias. With reference to the so called historical regression method used to evaluate nonrandomized cloud seeding operations one source of substantial bias was actually discovered [12]. This is the existence of several at least roughly identifiable types of storms, each with a different target-control regression line. The frequency of the particular types of storms varies from year to year. Thus, the normal target precipitation observed in a particular year may, and in some cases does, differ from the prediction based on a regression line calculated for a given epoch. Cases are on record [13] where a change in the epoch of reference causes a change in the conclusion regarding the effect of seeding: from a highly optimistic one to zero.

The existence of the types of storms with different regression lines has been verified twice. First this was done, somewhat grudgingly, by Thom ([14], p. 34). More recently this was done by Rapp and Schutz in an article [15] published by the National Academy of Sciences and National Research Council. The equivocations surrounding the admission of the existence of different types of storms and of their likely influence on the historical regression method present an interesting reading. The two authors state that they performed a "cursory" study (we agree) "to see whether there were *obvious* natural reasons" for the observed increases in target precipitation, other than seeding. Specifically, they performed a classification of storms by an approach about which: "It is admitted that the synoptic approach used here is not definitive (we agree), but is empirically sound" (we are not sure). After studying the regression lines for the two types of storms and comparing them with the seeded target precipitation the authors state: "We believe that there is a hint in these studies that a good portion of the deviation of the (seeded) precipitation from the regression line could be explained" (by the variability of types of storms). However, the final conclusion is that: ". . . it seems fair to postulate . . . that August 1962 does represent a month in which cloud stimulation did increase the precipitation . . ."

The more realistic conclusion is that the apparent excess precipitation of August 1962: (i) *may* be due to cloud seeding; (ii) that it may be due to the effect of storm types which a "cursory" study only indicated but which could have been established through a more "definitive" approach; and that also (iii) this apparent excess may be due to some causes of bias that are thus far unsuspected, or indeed (iv) to chance.

It will be remembered that prior to the work of the statistical Berkeley group, work conducted in cooperation with Mr. Edward Vernon, the existence of types of storms with different target-control regression lines, was not suspected. While we have mild regrets that the authorship of this discovery is not acknowledged, we do not contend that the type of storms is the only possible source of bias.

The role of randomization in the planning of an experiment is to eliminate the possibility of any unsuspected bias and to reduce the frequencies of erroneous judgments to levels that are both calculable and verifiable empirically. Naturally, mere inclusion of the randomization is no guarantee of soundness of an experiment. However, it is an important prerequisite.

A4. Statistical theoretical background

In the last section it was suggested that the planning of a future experiment with weather control be preceded by an examination of historical climatological data collected for the contemplated general area of the experiment. This examination, covering a substantial period of time, perhaps as much as a decade, would have the purpose of establishing the most advantageous elements of the design of the prospective experiment, such as the definition of the observational unit, the desirable predictor variables and the details of the target. The general method of procedure would be, essentially, by trial and error.

Substantive meteorological considerations would suggest several alternative possibilities of observational units (for example "storms" or periods of 24 hrs, and so forth), a target (its size, location, and whether fixed or adjustable) and some predictor variables (perhaps none or, alternatively, precipitation amounts in some comparison areas or, indeed, some nonprecipitation data). In other words, meteorological considerations would bring out several alternative designs for the planned experiment.

The next step in planning is to use the historical data in order to estimate (i) which of these designs would be the most effective if the future experiment were performed in conditions exactly similar to those reflected in the historical data, and (ii) what would the duration of the experiment have to be in order to insure that it will detect, with a reasonable degree of precision, what it is supposed to detect.

Naturally, the weather conditions of the historical period need not be duplicated during the future period contemplated for the experiment, and therefore, the precision of the future experiment will not be exactly as calculated. However, in this respect experiments with weather control are in no worse condition than, for example, experiments in medicine and in many other domains where one has to deal with uncontrollable observational units. In spite of the irregular and frequently serial variability of these units, progress in these domains is achieved customarily through projections of past experience into the future. Incidentally, this past experience indicates also the order of magnitude of errors to be expected in the projections.

The purpose of the present section is to introduce certain concepts and to indicate certain formulas relevant to the particular phase of planning an experiment just described. Numerical calculation based on these formulas are given in the next section. In both sections we shall pretend that a historical climatological study has been completed, providing data on precipitation in a contemplated target for some tentatively defined observational units, and also data for several contemplated predictors. The range of possibilities that a study of this kind may present is tremendous and in many cases the statistical treatment of the data is likely to require new methods. Also, an *a priori* treatment of the problem of projections may easily involve discussions of totally unrealistic situations. For these reasons further discussion will be concerned with real data and be limited to conditions reflected therein.

There is a notorious scarcity of easily available climatological data suitable for the present purposes. Essentially, we have at our disposal a single set provided by Project SCUD [3]. Here there were only 37 observational units, some of them with seeded and others with not seeded precipitation, much too few for a proper historical study. Nevertheless, we shall pretend that the hypothetical historical study of perhaps 370 units provided information coinciding with that in the 37 units of SCUD, all of which will be treated as if there were no seeding.

As described elsewhere by Marcella and John Wells [16] the observational data of Project SCUD, referring to the more interesting region *Ia*, have the following properties.

(i) Forecasting was so good that for each of the 37 observational units there was some rain in the target. Projecting this into the future, it will be assumed that the problem of the probability ϑ that there will be some rain in the target will not arise. The important problem relating to the frequent cases where a noticeable proportion of experimental units have zero precipitation is treated in a separate paper [17].

(ii) As is usually the case in our experience with nonzero precipitation, the precipitation amounts per observational unit appear to follow a Gamma distribution with density

$$p_Y(y|\gamma, \delta) = \frac{\delta^\gamma}{\Gamma(\gamma)} y^{\gamma-1} e^{-\delta y}. \tag{A4.1}$$

In the original paper, the precipitation amounts considered here are denoted by *RI*. We shall adopt the conventional notation Y and proceed on the assumption that in the future experiment Y will have the probability density (A4.1) with parameters γ and δ as estimated for SCUD.

(iii) The observations involved three predictor variables M, T and L. Also, it was found (see [16]) that the regressions of *RI* on M, T and L are approximately linear and that the conditional distribution of *RI* given M, T, L is approximately symmetric and has a constant variance. As is well known, in these conditions the assumption that the conditional distribution of *RI* given the predictors is normal, with a fixed variance, leads to tests that are "robust." This means that the deviations from the assumption of normality do not significantly affect the performance of the tests.

In our projection into the future we shall assume that there are some s predictors, say $X_1, X_2, \cdots, X_s$. Regarding the conditional distribution of target precipitation given the predictors, we shall consider two alternative situations. One of them coincides with what was found for SCUD, with Y, measured in inches, having a linear regression on the predictors and having, approximately, a normal conditional distribution with a fixed variance. The alternative assumption will be that the properties of linear regression, and of approximately normal conditional distribution with a fixed variance, are possessed not by the target precipitation Y measured in inches but, as is frequently the case, by a transformation of Y. More specifically, we shall assume that some rth root of Y possesses these properties. This may be the square root of Y or the cube root of Y, and so forth.

Our final assumption will be that, as expected by Spar, if there is an effect of seeding, this effect will be multiplicative. In more precise terms, this means that, with seeding, for any combination of values of the predictors $x_1, x_2, \cdots, x_s$, the expected value of the target precipitation is equal to that without seeding, multiplied by a factor independent of $x_1, x_2, \cdots, x_s$, all other properties of the conditional distribution of Y remaining unchanged.

In our calculations we found it convenient to represent this factor by θ so that θ represents a conventional measure of the effect of seeding. If $\theta = 1$ then seeding has no effect. If $\theta > 1$ then, on the average, seeding increases precipitation by a factor θ, and so forth.

The discussion in the earlier sections indicates the importance of being able to estimate the chance that the planned experiment will detect the effects that it is supposed to detect. This means the importance of being able to calculate the power of the statistical test that will be used.

For many statistical tests in frequent use, the calculation of power is still an unsolved problem. However, in a broad category of cases, including those reflected by the assumptions (i), (ii), and (iii), the experiment can be evaluated by any test of a special class, named $C(\alpha)$ tests [18], for which the asymptotic power can easily be calculated. The steps involved are as follows.

First the intended significance level α is arbitrarily selected. Next the normal deviate, say $\nu(\alpha)$, is read off from tables of the normal integral so as to satisfy the condition

$$\frac{1}{\sqrt{2\pi}} \int_{-\nu(\alpha)}^{+\nu(\alpha)} e^{-x^2/2}\, dx = 1 - \alpha. \tag{A4.2}$$

Then the asymptotic power of any $C(\alpha)$ test, the power corresponding to any preassigned θ is given by

$$\beta(\theta, \alpha) = 1 - \frac{1}{\sqrt{2\pi}} \int_{\tau-\nu(\alpha)}^{\tau+\nu(\alpha)} e^{-x^2/2}\, dx, \tag{A4.3}$$

where τ represents the so called noncentrality parameter, depending upon θ, upon the number N of experimental units, on the randomization, on the particular distributions of the observable variables, and, of course, on the particular $C(\alpha)$ test used. The calculation of τ will be discussed presently. Now we refer to figure A1 illustrating the relation between the level of significance α and the power $\beta(\theta, \alpha)$ prevailing for a fixed τ.

It is seen that, when τ is small, say $\tau = 0.5$, and if this value is calculated for the effects of seeding (as measured by θ) that are really important to detect, then the experiment planned is without value. Even if one adopts a level of significance as "liberal" as $\alpha = 0.1$, the chance of detecting the effect of seeding that is judged important will be only $\beta = 0.14$. In other words, even if we agree to assert the existence of an effect, when none exists with a frequency of one in ten, the existing important effect will be detected only in about one in seven cases. On the other hand, if we insist that β be of some more favorable value, say $\beta = 0.8$, we would have to put up with $\alpha = 0.77$, meaning that, on the

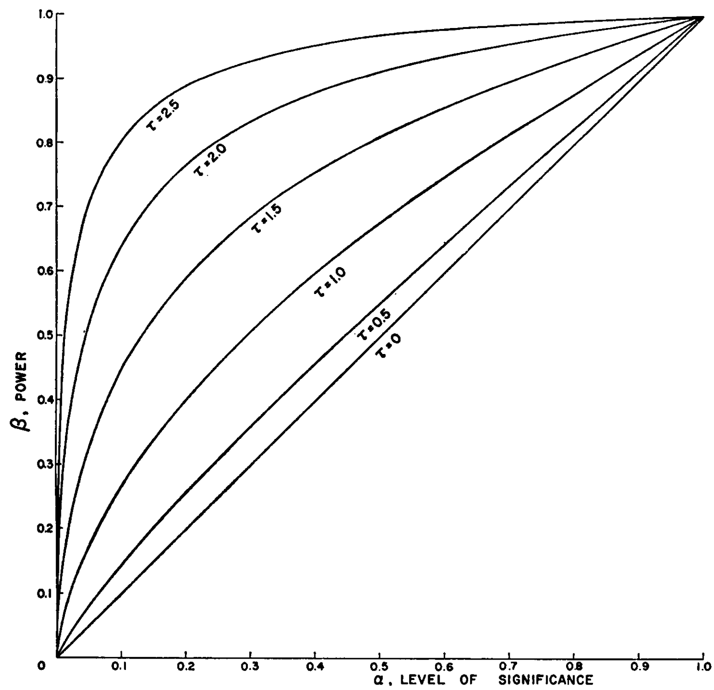

FIGURE A1

Relation between power and level of significance for various values of τ.

average, out of 100 similar experiments in which the treatment has no effect (thus $\theta = 1$), the test will lead to the assertion that an effect exists in about 77 cases!

Figure A1 indicates that anything like reasonable precision of the experiment is attained when τ is of the order of 2.5 or higher. With $\tau = 2.5$ and the level of significance $\alpha = 0.1$, the chance of detecting the hoped for effect would be about $\beta = 0.8$. With $\alpha = 0.05$ we shall have $\beta = 0.7$, and so forth. What is an acceptable chance of detecting what one wishes to detect is, of course, a subjective matter. In our opinion, for an important piece of experimentation, this chance should not be lower than about 0.8 and this is our reason for advocating some liberalism in selecting the level of significance.

The all important noncentrality parameter is given by the formula

$$\tau = \Delta[N\pi(1 - \pi)]^{1/2} \log_e \theta, \tag{A4.4}$$

which, in a rather attractive manner, relates τ to the several characteristics of the planned experiment, so that the role of each can be judged separately. Here, as explained earlier, θ is the factor multiplying the average normal precipitation in the target, a factor representing the hoped for effect of seeding *that is judged* *important to detect.* Thus, if it is judged important to detect the effect of seeding if it amounts to a 20 per cent or to a 40 per cent increase, then $\theta = 1.20$ or $\theta = 1.40$, respectively, and so forth. As formerly, N stands for the contemplated number of experimental units involved in the experiment. The symbol π refers to the randomization. If the experimental units are randomized in pairs, then $\pi = 1/2$. On the other hand, if unrestricted randomization is adopted, then π represents the probability that an experimental unit will be seeded. Here then, we come to a factor under control of the experimenter which influences the effectiveness of the experimental design and the optimal value of π maximizing the product $\pi(1 - \pi)$, and thereby τ, is $\pi = 1/2$. In some of the already performed experiments the value of π selected was $\pi = 2/3$, which resulted in a mild loss of precision of the experiment.

The most important factor in (A4.4) is that designated by Δ. The reason is that it is independent of either θ, N, or π and reflects all other characteristics of the experiment, namely the local conditions, the contemplated design and the statistical test used. While the local conditions, represented by the distributions of the observable random variables, are imposed on the experimenter, the other factors influencing Δ are at the experimenter's disposal and an effort to increase Δ is clearly indicated.

First, it is important to realize that, all other things being equal, the use of an inappropriate statistical test may reduce Δ to zero. The so called optimal $C(\alpha)$ test insures a value of Δ that is maximum compared to other tests of this class and, as shown by LeCam [19], this maximum frequently coincides with the absolute maximum. The formulas for the calculation of Δ for the optimal $C(\alpha)$ tests are as follows.

(1) If no predictor variables are used, and the target precipitation follows the Gamma distribution (A4.1), then

$$\Delta = \gamma^{1/2}. \tag{A4.5}$$

(2) If some s predictor variables $X_1, X_2, \cdots, X_s$ are used and if, as in SCUD, the target precipitation Y has a linear regression on the predictors

$$E(Y|x_1, x_2, \cdots, x_s) = a_0 + \sum_{j=1}^{s} a_j x_j, \tag{A4.6}$$

combined with approximate conditional normality and constant conditional variance, then

$$\Delta^2 = E\left[a_0 + \sum_{j=1}^{s} a_j X_j\right]^2 \Big/ \sigma^2. \tag{A4.7}$$

Here, the expectation is taken over the variability of the predictors, and the symbol σ represents the square root of the conditional variance of Y, given the predictors.

(3) Finally, if the properties of linearity of regression, as in (A4.6), of conditional normality and constant conditional variance are possessed not by the target precipitation itself but by its rth root, then

$$\Delta^2 = \left\{2 + \frac{1}{\sigma^2} E[a_0 + \textstyle\sum a_j X_j]^2\right\} \Big/ r^2, \tag{A4.8}$$

with an obvious change in the meaning of the symbols involved.

As mentioned, the $C(\alpha)$ tests are asymptotic tests and the conformity of real frequencies with those calculated from the above formulas can be expected when N is sufficiently large. Checks already performed indicate that with N of the order of 50 or more the performance of the tests satisfactorily agrees with predictions. For the method of deducing the above formulas, as well as for the details of criteria to be used in evaluations of the experiments, the reader is referred to [18].

All the above discussion refers to what might be called the mechanism of precision of an experiment. Turning to the very notion of precision, we wish to propose its measure, say N^*, defined as the minimum number of experimental units which insures that, with the adopted level of significance α, with optimal randomization $\pi = 1/2$, with the given distributions of the observable variables, and with the use of the optimal $C(\alpha)$ test, the effect of seeding θ that is judged important to detect will be detected with the preassigned probability β.

In order to obtain N^* it is sufficient to solve equation (A4.4) for N and to substitute in the right side the requisite values of τ (as determined by the preassigned values of α and β) and of the other parameters involved:

$$N^* = \frac{(2\tau/\log_e \theta)^2}{\Delta^2}. \tag{A4.9}$$

It will be noticed that the numerator in this formula depends only on α, β, and θ which the experimenter is at liberty to choose in conformity with his own opinions of the desired precision of the experiment. Contrary to this, the denominator Δ^2 depends on the conditions prevailing in the target area and on the design. It will be convenient to describe the chosen α, β, and θ as the desired "precision of the experiment." Then N^* might be described as the number of experimental units, or simply as the "size" of the experiment, insuring that, in conditions characterized by Δ^2 it will have the desired precision.

Tables AI and AII are intended to facilitate the computation of the numerator in (A4.9). Table AI gives directly $4\tau^2$ corresponding to several combinations of

TABLE AI

VALUES OF $4\tau^2$ IN TERMS OF α AND β

β \ α	0.10	0.05	0.02	0.01
0.70	18.82	24.69	32.51	38.45
0.80	24.73	31.40	40.14	46.72
0.90	34.26	42.03	52.07	59.52
0.95	43.29	51.98	63.08	71.26

α and β. Table AII gives $1/2 \log_e \theta$ as a function of the percentage change in precipitation that the experimenter might consider "important to detect."

TABLE AII

$1/2 \log_e \theta$ IN TERMS OF PERCENTAGE CHANGE IN MEAN RAIN DUE TO SEEDING

Per cent change	−80	−40	−20	−10	+10	+20	+40	+80
$1/2 \log_e \theta$	0.386	3.83	20.08	90.08	110.1	30.08	8.83	2.89

It will be noticed that a given percentage increase in rain due to seeding is more difficult to detect than the same percentage decrease.

A5. Number of experimental units required to attain the preassigned precision of a planned experiment

In this section we perform numerical calculations leading to estimates of the size N^* of an experiment insuring a preassigned precision. In other words, we shall estimate the number N^* of experimental units insuring a preassigned probability β that an effect of seeding, which is considered important, will be found significant at a preassigned level α. As indicated earlier, this will be done under the assumption that the planned experiment will be performed in physical conditions coinciding exactly with the estimates obtained for all the 37 units in SCUD data and ignoring the fact that some of them were seeded. Considerations of alternative designs of the planned experiment will be limited to (i) the possibility of using no predictor variables and (ii) the possibility of using one or more predictor variables M, T and L, which were actually used in Project SCUD.

As mentioned before, the exact definition of "desired precision" is subjective. Our own choice is

(A5.1) $$\alpha = 0.1, \qquad \beta = 0.9, \qquad \text{and} \qquad \theta = 1.4.$$

In other words, the calculations are performed to determine the minimum size N^* of an experiment in which the chance of finding a 40 per cent increase in rain significant at 10 per cent is equal to 0.9. Using tables AI and AII, the numerator of formula (A4.9) is found to be $2\tau^2/\log_e \theta = 34.26 \times 8.83 = 302.5$. Thus, for each of the possibilities contemplated we shall have to compute

(A5.2) $$N^* = \frac{302.5}{\Delta^2}.$$

We begin by considering the possibility (i) of using no predictor variables in the planned experiment. The formula for computing Δ^2 appropriate for the situation considered is (A4.5). It was found that the RI precipitation data of Project SCUD are satisfactorily fitted by the Gamma distribution with the shape parameter $\gamma = 1.70$. Referring to (A4.5) and substituting in (A5.2), we found

$$N^* = \frac{302.5}{1.70} = 177.9, \tag{A5.3}$$

say 178. Taking into account that it required two years to accumulate 37 observations, $N^* = 178$ means about 10 years.

At this point it is appropriate to mention that, probably due to the immense, 1000 mile long, target and to the particular type of storms studied in SCUD, the corresponding value of $\gamma = 1.70$ is unusually large. Table AIII gives a sample of values of γ as estimated for several sets of data.

TABLE AIII

SAMPLE OF ESTIMATES OF γ

Source of Data	No. of Observations	No. of Observations per Season	$\gamma \pm$ S.E.
SCUD	37	19	1.70 ± 0.36
Grossversuch IIIa	105	15	0.78 ± 0.09
Arizona I	89	22	0.66 ± 0.08
Grossversuch IIIb	242	35	0.54 ± 0.04
Little Egypt, Ill.	231	46	0.44 ± 0.03
East Central, Ill.	377	38	0.39 ± 0.02

The values given in table AIII are the maximum likelihood estimates obtained for each of the six sets of data. For Grossversuch III many such estimates were obtained corresponding to different subtargets and to different categories of experimental units. The two values given in the table are the extremes of those based on at least 100 observations. The estimate obtained for the first of the two Arizona experiments (for which the data are now available) happens to be exactly in the middle between these two extremes. The last two lines in the table refer to two sets of data kindly furnished to us by Mr. S. A. Changnon, Jr. [20]. They represent average precipitation amounts in many gages in two distinct areas in the State of Illinois, one labeled Little Egypt and the other East Central.

It is seen that the estimated shape parameters fall, roughly, into three groups. The top value corresponds to SCUD. For Grossversuch III and for Arizona the typical value appears to be 0.66. For the two targets in Illinois, γ is about 0.4. If the planned experiment were to be performed without predictors either in conditions of Arizona or of Grossversuch III, or in conditions of Illinois, the requisite number of experimental units would have to be more than double that computed for SCUD, in one case, and more than four times as large as for SCUD, in the other. However, in judging these numbers it is important to take account of the number of experimental units of a given kind that one may expect per season.

Leaving aside SCUD because of its nontypical target, it is interesting to speculate about the reasons for the difference between the conditions in the Arizona experiment and the Grossversuch III, on one hand, and in the two areas in Illinois, on the other. The two categories of conditions differ in the definition of the experimental unit. In Arizona and in Grossversuch III the experimental units were of fixed duration. On the other hand, the experimental units chosen in Illinois were storms, some of which lasted more than two days while some others were of very short duration so that during one single day there may have been as many as three different storms, each counted as a separate observational unit. Referring to formula (A4.1), it will be seen that the smaller the parameter $\gamma > 0$, the more frequent must be experimental units with minute amounts of target precipitation. This suggests that the smallness of γ for the Illinois data may be due to the particular choice of the experimental unit. It is possible that, if the "storm" is replaced, say, by a properly chosen 12 hour period, the value of γ would be increased with an increase in the precision of the experiment. On the other hand, this change in the experimental unit may well decrease the number of experimental units per season, which may increase the requisite duration of the experiment. The choice of the experimental unit must be done after appropriate examination of the historical data leading to a complete summary evaluation of all the factors in their conflicting influences on the duration of the experiment.

The result (A5.3) and the data in table AIII indicate that, unless one is prepared to relax considerably the definition of the desirable precision of the experiment, now characterized by $\alpha = 0.1$, $\beta = 0.9$ and by considering that a 40 per cent increase in precipitation is important to detect, rain stimulation experiments without predictor variables must be of distressingly long duration. In conditions of SCUD this would be about a decade. In conditions of Arizona about two decades would be needed, and so forth. Accordingly, we now turn to the examination of the effects of predictor variables. This means the evaluation of formula (A5.2) with Δ^2 evaluated from (A4.7).

The numerator in this formula, say $Y^2(x)$ may be estimated by averaging, over all the experimental units in the hypothetical historical study, the square of the ordinate of the regression plane (or line, or hyperplane) of the target precipitation on the predictor variables contemplated. The denominator is, simply, the corresponding residual variance. For Project SCUD the relevant calculations are given in [1]. Using these results, and considering in turn the possibilities of using only one predictor variable, either M or T or L, any two of them, or all three, estimates of the required size of the planned experiment were obtained. All the results, including some intermediate stages, are summarized in table AIV.

It is seen that the inclusion of the predictor variables has an effect on the precision of the experiment that cannot be described with less emphasis than dramatic. In fact, the use of all three predictors, and for that matter, the use of only two of them, M and T, reduces the required size of the experiment from 178 units to only 42, by a factor of 4.2! In considering this result, it is important

TABLE AIV

REQUIRED SIZE N^* OF THE EXPERIMENT IN RELATION TO PREDICTOR VARIABLES USED

Predictors Used	Mean $Y^2(x)$	Estimated Residual Variance	Required Size N^*
none	—	—	178
M	0.113	0.029	78
T	0.119	0.022	57
L	0.106	0.036	102
M, T	0.125	0.017	42
T, L	0.121	0.022	56
L, M	0.116	0.027	71
M, T, L	0.125	0.018	42

to remember that, while the numerical estimates of N^* in the last column depend on the definition of the desired precision of the experiment, that is, on $\alpha = 0.1$, $\beta = 0.9$ and $\theta = 1.4$, ratios like $178/42 = 4.2$ are independent of this definition. Thus, in conditions coinciding with those of Project SCUD, the inclusion of the predictors M and T would decrease the required size of the experiment by a factor of 4.2, irrespective of what one chooses to consider as a satisfactory precision of the experiment.

Table AIV indicates that the effectiveness of the three predictors M, T, and L is very unequal. The best predictor is undoubtedly T. The least satisfactory is L, so much so that the addition of L to the two other predictors M and T does not increase the precision of the experiment.

The results of table AIV appeared to the present authors as a most pleasant surprise. From past experience we became accustomed to the idea that precipitation amounts in some two not very distant areas are correlated, which is the basis for the use of control areas. However, the variables M, T, and L are not precipitation amounts but are functions of observations concerned with weather conditions that to a layman may seem remote. The variable M is a somewhat elaborate measure of the east-west pressure gradient. The variable T is described by Spar as the influx of water vapor into the target area, and L is, essentially, the latitude of the cyclone center estimated for the zero hour. With these definitions one source of our surprise is the effectiveness of M and T. The other source of surprise is that, for given M and T, the value of L is irrelevant.

It would be most satisfactory if the effectiveness of predictors invented for SCUD could be immediately generalized for other experiments, organized in other localities and with much smaller targets. Regretfully, because of the disparity between the target area in SCUD and the areas of targets that have now become customary, no such generalization is justifiable without prior empirical verification. As things stand now, the only general conclusions that the above analysis suggests are: (i) with relatively brief units of observations and without the use of any predictor variables, in order to satisfy modest requirements of precision, experiments with rain stimulation must last a discouragingly long time; and (ii) in some conditions at least, the inclusion of properly selected collateral observations, yielding one or more predictors, may increase the precision of the experiment dramatically.

Before concluding this section, the following two remarks are in order. The first remark is that the results just discussed, interesting as they are, are subject to considerable uncertainty. This is due to the very small number, only 37, of observations in SCUD, which underlie all the calculations exhibited in table AIV. Our attitude to these results would have been quite different if they were based on a real historical study covering perhaps 400 experimental units. In particular, it is quite possible that, as a result of such study, it would appear that the predictor L is not valueless.

Our second remark relates to the validity of calculations based on historical data leading to results such as those exhibited in table AIV. This question of validity splits into two subquestions. One is whether the calculated N^* really insures that in a long series of experiments conducted *strictly in conditions revealed by the historical study*, the application of the test at the preassigned level of significance α will detect the effect with frequency β if this effect corresponds to the chosen θ. This is a very important question which, particularly because of the asymptotic character of the optimal $C(\alpha)$ tests, should be asked and should be answered empirically. Following the lines which led to figure 6 in the main body of this paper, or the lines of Changnon and Huff, this is not a difficult question. As a general rule, the frequencies of errors resulting from the application of the optimal $C(\alpha)$ tests agree reasonably with expectation even when the number of observations is only of the order of a few tens.

The other subquestion is whether and to what extent the conditions revealed by the historical study are likely to be duplicated during the time period of the future experiment. No definitive answer to this question can be given. All that is possible to do is to use the data of a rather long historical study to determine the variability of the estimate of N^* if it is based on a single year, on two consecutive years, on three, and so forth.

Because of the lack of data involving predictor variables and referring to several years in the past, the illustration of the above study is only possible for the case of no predictor variables. Figure A2 gives the variability of estimates of N^* based on varying numbers of consecutive years calculated from Grossversuch III data for zone III.

It is seen from figure A2 that estimates of N^* based on just one year of observations vary in very broad limits, from about 350 to about 550 units of observation. With a six year period the variation is reduced very markedly. However, it must be remembered that in this case there are only two estimates of N^* and that, by their structure, they are very strongly correlated.

Of the six graphs in figure A2, the first two are substantially more interesting than the others. In the first the consecutive values of N^* are estimated each from a separate set of data, without overlap. In the second graph, there is an

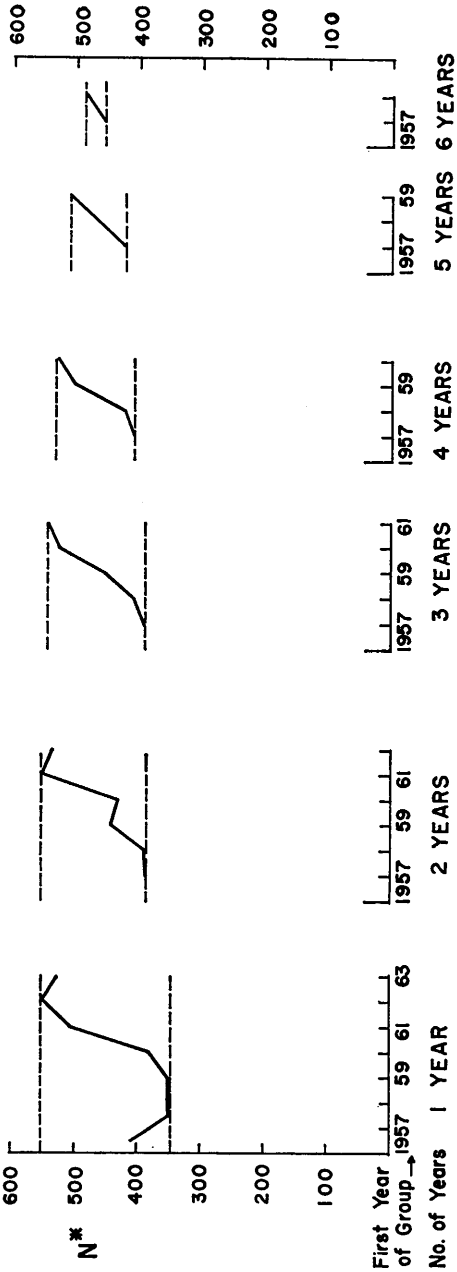

FIGURE A2

Estimates of N^* based on varying numbers of years of observation, using data of Swiss experiment [4].

overlap between the first estimate and the second, but not the third, and so forth. Both graphs illustrate the striking phenomenon of what may be called serial variability in the weather pattern which must be taken into account in weather control experimentation. In fact, the first graph suggests something like a periodicity of about ten years cycle (sunspots?). If any such periodicity is detected, then it could perhaps be used to tighten the estimates of N^*.

As mentioned, similar calculations could and should be done using predictors. However, here the situation is likely to be a little more favorable than figure A2 suggests. The point is that the serial variability exhibited in figure A2 would affect not only the target precipitation, but also the predictor variables which would decrease the variability of estimated N^*. Also, our experience indicates that the correlation between target precipitation and that in control areas seems to depend on identifiable types of storms. Of two comparison areas C_1 and C_2, the first appears to correlate strongly with the target with storms of one type, when C_2 is ineffective, and vice versa.

Before concluding we wish to express our hearty appreciation to Miss Kang Ling, Mrs. Jeanne Lovasich, and Mrs. Marcella Wells for the infinite patience and care in performing the many computations that underly the numerical results given in this paper.

REFERENCES

[1] R. A. Fisher, *The Design of Experiments*, London, Oliver and Boyd, 1936 (1st ed.).
[2] R. A. Fisher and Frank Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, London, Oliver and Boyd, 1943 (4th ed.).
[3] J. Spar, "Project SCUD," *Meteor. Monogr.*, Vol. 2 (1957), pp. 5–23.
[4] P. Schmid, "On 'Grossversuch III,' a randomized hail suppression experiment in Switzerland," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1967, Vol. 5, pp. 141–159.
[5] T. J. Henderson, "Tracking silver iodide nuclei under orographic influence," paper prepared for the *24th National Meeting, American Meteorological Society*, Reno, Nevada, 1965.
[6] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. Roy. Soc. Lond. Ser. A*, Vol. 231 (1933), pp. 289–338.
[7] R. A. Fisher, "Theory of statistical estimation," *Proc. Cambridge Phil. Soc.*, Vol. 22 (1925), pp. 700–725.
[8] J. Neyman, "Outline of a theory of statistical estimation based on the classical theory of probability," *Phil. Trans. Roy. Soc. Lond. Ser. A*, Vol. 236 (1937), pp. 333–380.
[9] E. L. Lehmann, *Testing Statistical Hypotheses*, New York, Wiley, 1959.
[10] S. A. Changnon, Jr. and F. A. Huff, "The effect of natural rainfall variability in verification of rain modification experiments," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1967, Vol. 5, pp. 177–198.
[11] M. Twain, *A Connecticut Yankee in King Arthur's Court*, New York, Webster, 1889.
[12] J. Neyman and E. L. Scott, "Further comments on the 'Final Report of the Advisory Committee on Weather Control'," *J. Amer. Statist. Assoc.*, Vol. 56 (1961), pp. 580–600.
[13] G. W. Brier and I. Enger, "An analysis of the results of the 1951 cloud seeding operations in Central Arizona," *Bull. Amer. Meteor. Soc.*, Vol. 33 (1952), pp. 208–210.

[14] H. C. S. THOM, "An evaluation of a series of orographic cloud seeding operations," *Final Report of the Advisory Committee on Weather Control*, Washington, 1957, Vol. 2, pp. 25–50.

[15] R. RAPP and C. SCHUTZ, "Preliminary study of meteorological effects on specific seeding trials," *Weather and Climate Modification Problems and Prospects*, National Academy of Science-National Research Council, 1966, Vol. 2, pp. 169–173.

[16] J. WELLS and M. A. WELLS, "Note on Project SCUD," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1967, Vol. 5, pp. 357–369.

[17] J. NEYMAN and E. L. SCOTT, "Note on techniques of evaluation of single rain stimulation experiments," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1967, Vol. 5, pp. 371–384.

[18] J. NEYMAN and E. L. SCOTT, "Asymptotically optimal tests of composite hypotheses for randomized experiments with noncontrolled predictor variables," *J. Amer. Statist. Assoc.*, Vol. 60 (1965), pp. 699–721.

[19] L. LECAM, "On the asymptotic theory of estimation and testing hypotheses," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1956, Vol. 1, pp. 129–156.

[20] S. A. CHANGNON, JR., Personal communication.