

Frequentist Probability and Frequentist Statistics

Author(s): J. Neyman

Source: *Synthese*, Vol. 36, No. 1, Foundations of Probability and Statistics, Part I (Sep., 1977), pp. 97-131

Published by: Springer

Stable URL: <http://www.jstor.org/stable/20115217>

Accessed: 25-05-2016 23:52 UTC

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
<http://about.jstor.org/terms>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



Springer is collaborating with JSTOR to digitize, preserve and extend access to *Synthese*

J. NEYMAN

FREQUENTIST PROBABILITY
AND
FREQUENTIST STATISTICS*

I. INTRODUCTION

The present paper is prompted by two stimuli. One is the gracious invitation of Professor Jaakko Hintikka to contribute to the issue of his journal especially given to foundations of probability and statistics. The other stimulus is multiple: letters from friends calling my attention to a dispute in journal articles, in letters to editors, and in books, about what is described as ‘the Neyman-Pearson school’ and particularly what is described as Neyman’s ‘radical’ objectivism. While being grateful to my friends for their effort to keep me informed, I have to admit that, owing to a variety of present research preoccupations, I have not read the whole of the literature mentioned to me. However, I glanced at the published exchange of letters and at the book by de Finetti [1]. My reactions are somewhat mixed. First, I feel honored by the attention given to my writings, primarily those published more than a quarter of a century ago (see [2]). Next, I feel a degree of amusement when reading an exchange between an authority in ‘subjectivistic statistics’ and a practicing statistician, more or less to this effect:

The Authority: ‘You must not use confidence intervals; they are discredited!’

Practicing Statistician: ‘I use confidence intervals because they correspond exactly to certain needs of applied work.’

My third present sensation is that of surprise at the intensity of feeling apparent in some publications. One illustration is the following statement by de Finetti and L. J. Savage, quoted from [1], page 192. Here, italics are mine:

Every progress in scientific thought involves struggle against the distortion of outlook generated by the habits of ordinary language, the deficiencies and captiousness of which

Synthese 36 (1977) 97–131. All Rights Reserved.

Copyright © 1977 by D. Reidel Publishing Company, Dordrecht, Holland.

must be denounced, combatted, and overcome, case by case. This has been done, and is being done, explicitly by all those scientists who find themselves in need of introducing and defending new concepts. And the struggle is rendered harder by the support that anyone averse to novelty finds in the opposing tendency of rhetoricians to identify thought and reasoning accidental with details – generally unfortunate or, at any rate, obsolete – of ways of speaking *invented for the conditions of thousands of years ago.*

When I read this statement the scene of Giordano Bruno being led to the stake comes to my mind.

The above passage is quoted from de Finetti's section, 'Critical Examination of Controversial Aspects' and is not necessarily addressed to my views or writings, to be 'denounced' or 'combatted'. My views are discussed in the two preceding sections, dealing with 'The Rise of Objectivistic Concepts' and with 'The Erosion of Objectivistic Positions'. It is here that I am quoted and, occasionally, misrepresented. Two quotes from page 175 must suffice.

Other objectivistic works suggest rules for guiding the choice, but quite untenable ones, such as the minimax rule or else holding α to some fixed value (like 1% or 5%) regardless of β .

Many weaknesses of the objectivistic theory that can be detected and analysed in such perfectly coherent formulations as Neyman's, like the one just alluded to, result from the artificiality with which the procedures have to be constructed and from the impossibility of utilizing all the information . . .

The purpose of the present paper is to outline, briefly and non-technically, my views on what I like to call the *frequentist* theory of probability and the *frequentist* theory of statistics, and to illustrate their applications in a few domains of study of nature in which I have been involved. My term 'frequentist' seems to correspond to what de Finetti labels 'objectivistic', but there is a difference which I hope the following pages will clarify.

II. THE PHENOMENON OF APPARENTLY STABLE RELATIVE FREQUENCIES AS THE SOURCE OF THE FREQUENTIST THEORIES OF PROBABILITY AND OF STATISTICS

1. *General Ideas*

As described in a recent paper [3], it is my opinion that, directly or indirectly, all mathematical disciplines stem from human efforts to study Nature. Particularly this applies to 'young' mathematical disciplines or,

more precisely, to the relatively early period of their development. Later, the discipline reaches a state of maturity and begins to live its own life. Contacts with substantive studies diminish and recede, and new theoretical developments are motivated by their own intrinsic interest. However, cases of 'feedback' do occur from time to time. This happens when a 'practitioner' in a mathematical discipline happens to develop interest in some domain of study of Nature. As I see it, this is just the case of mathematical statistics in its present state of development. My own status is that of a practitioner in the theory of statistics with deep interest in 'chance mechanisms' that operate in Nature.

In a humorous vein we might say that the honor of discovering the category of natural phenomena that generated the frequentist theory of probability belongs to the first crook who loaded his dice. Before embarking on this project the particular individual must have realized that the relative frequencies of a die falling this way or that way are 'persistent' and constitute this die's measurable properties, comparable to its size and weight. Having discovered this fact (and this was a 'scientific discovery'), the crook decided to use the discovery for his own benefit (and this might be described as the initiation of a special 'technology').

It so happens, see [3], that very substantial sections of modern science and of technology are working hard more or less to follow the steps of the above crook (no offense is intended!).

2. *Three Steps*

There are three distinct steps in this process:

(i) Empirical establishment of apparently stable long-run relative frequencies (or 'frequencies' for short) of events judged interesting, as they develop in nature.

(ii) Guessing and then verifying the 'chance mechanism', the repeated operation of which produces the observed frequencies. This is a problem of 'frequentist probability theory'. Occasionally, this step is labeled 'model building'. Naturally, the guessed chance mechanism is hypothetical.

(iii) Using the hypothetical chance mechanism of the phenomenon studied to *deduce* rules of adjusting our actions (or 'decisions') to the observations so as to ensure the highest 'measure' of 'success'. Naturally,

the definition of 'success' is a subjective matter, outside of mathematics. So is the 'measure' of success. On the other hand, the *deduction* of the 'rules of adjusting our actions' is a problem of mathematics, specifically of mathematical statistics.

Incidentally, the early term I introduced to designate the process of adjusting our actions to observations is 'inductive behavior'. It was meant to contrast with the term 'inductive reasoning' which R. A. Fisher used in connection with his 'new measure of confidence or diffidence' represented by the likelihood function and with 'fiducial argument'. Both these concepts or principles are foreign to me.

I do deal with likelihood function and, occasionally, calculate the maximum likelihood estimators. However, I do so not as a matter of principle, but only in those cases when the frequency properties of the estimators fit my purposes. In other cases, illustrated in [4], other estimators appear preferable.

As to 'fiducial argument', a detailed analysis of what was published in the 1930's showed (see pp. 375–393 in [2]) that it is a conglomeration of mutually inconsistent assertions, not a mathematical theory.

III. ILLUSTRATIONS OF THE THREE STEPS

3. *The Crook and Insurance*

Presumably, after loading his die the crook must have tossed it quite a few times in order to acquire information about how frequently it falls on each of its sides. Without such information he could not hope to derive much profit from the loading.

Similarly, in order to conduct their business properly, the insurance companies must use the actuarial tables of mortality, of sickness, of accidents, etc.

4. *Models of Chance Mechanisms Operating In Nature*

Several modern books on probability [5, 6, 7] include chapters or sections with telling titles, such as the following: 'Birth and Death Processes', 'Queing Process', 'Branching Process', etc. Each such title refers to a hypothetical chance mechanism defined and developed with the idea that

its operation might mimic the natural phenomenon indicated in the title. The term frequently used to designate the hypothetical chance mechanism is 'stochastic model'. Another word frequently used in connection with chance mechanisms is 'random', e.g. 'random events', 'random variable', etc.

As things stand now, probability theory is a fully mature mathematical discipline 'living its own life', and this includes the three books quoted. The origins of this theory appear to be in the famous monograph of Kolmogorov [8]. Later developments may be symbolised by the books of Doob [9], of Dynkin [10], and of Loève [11]. The ties with empirical frequencies are indicated by Kolmogorov. These ties are more pronounced in the writings of von Mises, who built a probability theory of his own. While I prefer the theory of Kolmogorov, I am appreciative of von Mises' efforts to separate a frequentist probability theory from the intuitive feelings of what is likely or is unlikely to happen. The following somewhat emphatic quotation is from his book [12].

Mit der Frage, ob und wie wahrscheinlich es ist, dass Deutschland noch einmal Krieg mit der Republik Liberia führen wird, hat unsere Wahrscheinlichkeitstheorie nicht das mindeste zu tun.

Ordinarily, the 'verification', or 'validation' of a guessed model consists in deducing some of its frequentist consequences in situations not previously studied empirically, and then in performing appropriate experiments to see whether their results are consistent with predictions. Very generally, the first attempt at verification is negative: the observed frequencies of the various outcomes of the experiment disagree with the model. However, on some lucky occasions there is a reasonable agreement and one feels the satisfaction of having 'understood' the phenomenon, at least in some general way. Later on, invariably, new empirical findings appear, indicating the inadequacy of the original model and demanding its abandonment or modification. And this is the history of science!

An outstanding example is the history of genetics, beginning with Mendel's laws of inheritance. They are 'frequentist'. Discovered in the mid-nineteenth century, overlooked or ignored for a long time, these

laws were rediscovered at the turn of the century, and were seemingly 'confirmed'. Then some further studies revealed a number of unanticipated details, such as 'linkages' between genes, 'mutations', etc., all reflected in a variety of 'frequentist' findings. In consequence, the original chance mechanism invented by Mendel, fundamental as it continues to be, is now far away removed from the attention of scientists concerned.

The idea of frequentist models of natural phenomena seems to be due to Émile Borel. In fact, in his book [13], first published in 1909, Borel identified the construction of stochastic models with the general problem of mathematical statistics:

Le problème général de la statistique mathématique est le suivant. Déterminer un système de tirages effectués dans urnes de composition fixe de telle manière que les résultats d'une série de tirages, interprétés à l'aide de coefficients fixes convenablement choisis, puissent avec une très grande vraisemblance conduire à un tableau identique au tableau des observations.

Here, the 'properly selected coefficients' appear to designate estimates of the parameters involved in the model, the values of which were left unspecified.

5. *Inductive Behavior*

Problems of category (iii) may be exemplified by efforts to 'validate' or to 'verify' the Mendelian laws. Briefly and roughly, the situation may be summed up by the following question: shall we conduct our studies of heredity on the assumption that Mendelian laws may be realistic, or shall we ignore them? Early in this century many experiments were performed intended to answer this question. Let us consider the essence of such an experiment. Certain organisms such as plants or animals, are cross-fertilized producing n progeny. The Mendelian laws predict that this progeny would fall into a number, say s , of distinct categories and that the i th category has the specified probability p_i , for $i = 1, 2, \dots, s$. The performed experiment results in n_i individuals of the i th category. The 'expected value' of n_i is the product np_i . But the Mendel law specifies only a chance mechanism of inheritance and it is not contended that each n_i must be equal to its expectation. The question is about an intelligible

methodology for deciding whether the observed numbers n_i , differing from np_i , contradict the stochastic model of Mendel.

Early in this century this question was the subject of a lively discussion by Borel [14] and others. Borel was optimistic but insisted that: (a) the criterion to test a hypothesis (a 'statistical hypothesis') using some observations must be selected *not after the examination of the results of observation*, but before, and (b) this criterion should be a function of the observations 'en quelque sorte remarquable'. It is these remarks of Borel that served as an inspiration to Egon S. Pearson and myself in our effort to build a frequentist theory of testing hypotheses.

IV. SKETCH OF THE THEORY OF TESTING STATISTICAL HYPOTHESES

6. *Basic Concepts*

Regretfully, I must admit that E.S.P. and I were rather slow. Our first, rather long but inadequate attempt [15] was published in 1928, and the paper with some really new ideas [16] appeared in 1933, five years later. This illustrates the observation that the most difficult parts of mathematical research in any field consist (a) in noticing the existence of an important problem (in the present case this was done by Borel in 1909, but we learned about it in the 1920's), and (b) in formulating the problem so it makes mathematical sense, and to initiate its solution.

The basic ideas of the theory initiated in [16] are (i) the existence of two kinds of errors possible to commit while testing a hypothesis, (ii) the notion that these two kinds of error may be of unequal practical importance, (iii) that a desirable method of testing hypotheses must ensure an acceptably low probability, say α , of the more important error, and (iv) that, point (iii) being satisfied with an acceptable α , the probability of the less important error should be minimized.

Remark. The reader will realise that the above brief description of the basic problem of testing statistical hypotheses is necessarily somewhat rough. In particular, the concepts of 'acceptable low probability level α ' and of 'minimizing the probability of the less important error' are described with an effort to avoid a number of technicalities, some of them quite difficult. Here an interested reader is referred to [18]. The modern

version of the theory is available in a number of books, frequently with many editions and translations. See [19, 20, 21].

Now it is necessary to introduce some terminology. We must begin with general considerations. Whatever hypothesis, say H , may come under consideration (for example the hypothesis of validity of Mendel laws), an attempt to test it using some experimental data implies the admission that the hypothesis H may be false. In turn, this implies that, in addition to H there must exist some other hypotheses, one of which may conceivably be true. Here, then, we come to the concept of the 'set of all admissible hypotheses' which is frequently denoted by the letter Ω . Naturally, Ω must contain H . Let \bar{H} denote the complement, say $\Omega - H = \bar{H}$. It will be noticed that when speaking of a test of the hypothesis H , we really speak of its test 'against the alternative \bar{H} '. This is quite important. The fact is that, unless the alternative \bar{H} is specified, the problem of an optimal test of H is indeterminate [18]. In the 1930's this was one of the subjects of dispute with R. A. Fisher.

As mentioned at the outset, a test of the hypothesis H can lead to an error which may be of two kinds. One kind of error consists in the rejection of H when it is true. The other kind of error is committed when H is false (and, therefore, \bar{H} is true) and we fail to reject H . According to circumstances and according to the subjective attitudes of the research worker, one of these two kinds of error may appear more important to avoid than the other.

In this connection, modifying somewhat the notions explained in [16], we introduce two important terms: (a) the error which is the more important to avoid will be called 'error of the first kind', and (b) of the two alternatives H and \bar{H} , the one the unjust rejection of which constitutes the error of the first kind, will be called 'the hypothesis tested'. With this convention, the error of rejecting the hypothesis tested when true will mean the 'error of the first kind', the one we judge to be the more important error to avoid. As mentioned, the aim of the theory of testing is to reduce the probability of this error to some acceptable low level α . This level is called 'significance level'.

Now we introduce the term 'power of a test'. This term means the probability of avoiding the error of the second kind. Customarily, it is denoted by β . The word 'power' connotes 'power of detecting the falsehood of the hypothesis tested'.

E.S.P.'s and my effort at a theory of testing statistical hypotheses [16] had a double aim: first reduce the probability of the first kind of error to a preassigned level α and, when a class of tests all satisfying this condition is found, determine the one with greatest possible power. This would be the 'optimal' test.

Remark. The above description of the aim of the theory of testing hypotheses is intentionally oversimplified. Otherwise, the present article would have to be both too long and too technical. One kind of technical difficulty occurs when it is found that the originally defined 'optimal test' fails to exist in some interesting category of cases [22]. Then one looks for a 'compromise optimal test'. One example of this kind is the concept of a 'minimax' test introduced by E.S.P. and myself [23], which de Finetti finds 'quite untenable' (see one of the quotes above). The word 'untenable' suggests the situation in which either E.S.P. or I made efforts to enforce the minimax procedures as a matter of principle or dogma, or at least to 'sell' them to some 'consumers', such as 'practicing statisticians'. We do not. Our object was, and continues to be, to investigate test procedures within the frequentist theory possible to be applied in a variety of situations. The 'minimax' procedure was invented for a particular case in which (a) the unavoidable errors are to be 'paid for' causing 'losses' to the practitioner, and (b) in which it may be desired to 'minimize the maximum' possible loss. Incidentally, the term 'minimax' is not ours. This term was introduced by Abraham Wald [24], a great talent who perished in an airplane accident in 1950. He unified and generalized all the earlier efforts at developing the mathematical theory of statistics. In fact, the appearance of Wald's works may be considered as marking the 'maturity' of mathematical statistics as an independent mathematical discipline. Wald's work is discussed by de Finetti in his section, 'The Erosion of Objectivistic Positions'. This title illustrates a difference in the attitudes of de Finetti and myself.

A more recent and a very interesting compromise concept of optimality of a test is due to Robert Davies [25].

Now, we need two more technical terms: 'simple' and 'composite' hypotheses. A hypothesis H is called 'simple' if it specifies completely the probability distribution of the observable variables. Otherwise it is called 'composite'.

Here are a few examples illustrating the concepts introduced.

7. *Simple and Composite Hypotheses*

The hypothesis H_1 that a given coin is 'fair' is simple. When tossed, this coin can fall in one of two ways: 'heads' or 'tails'. The hypothesis H_1 means that both outcomes of a toss have the same probability, namely one-half.

This is contrasted with the alternative, say H_2 , which asserts ONLY that the coin is 'biased' and nothing else. Thus, according to H_2 , the probability of the coin falling 'heads' is not equal to one-half and may be any other number between zero and unity. It follows that H_2 is a 'composite hypothesis'. Obviously, this composite hypothesis H_2 represents a combination (or is 'composed') of an infinity of simple hypotheses, such as that the probability of 'heads' is $\frac{1}{3}$, or $\frac{3}{4}$, etc. etc.

8. *Hypothesis Tested*

Activities of the U.S. Food and Drug Administration include the testing of chemicals to be sold to customers as food additives or cosmetics. One of the questions asked is whether these chemicals are carcinogenic. Consider an experiment with mice (hoping that its results will also apply to man) intended to determine whether a chemical A is carcinogenic or not. This experiment, with m mice exposed to A and n control mice, will show some numbers X and Y of mice which died from cancer.

Our question is: What is our 'hypothesis tested'? To answer this question we must first answer another question: which error in testing is the more important to avoid?

As usual, there are two possible errors. The verdicts about A may be: (i) ' A is carcinogenic', and (ii) ' A is not carcinogenic'. Each of these verdicts may be wrong. Which of these errors is our 'error of the first kind'? Here we come to the subjectivity of judging importance. From the point of view of the manufacturer the error in asserting the carcinogenicity of A is (or may be) more important to avoid than the error in asserting that A is harmless. Thus, for the manufacturers of A , the 'hypothesis tested' may well be: ' A is not carcinogenic'. On the other hand, for the prospective user of the chemical A the hypothesis tested will be unambiguously: ' A is carcinogenic'. In fact, this user is likely to hope that the probability of error in rejecting this hypothesis be reduced to a very small value!

9. *Importance of the Power of the Test Used*

The mathematical results developed in [16], followed by those of many other workers, provide the possibility of reducing the probability of the first kind of error, at least approximately, to a preassigned level of significance α , applicable in a great variety of situations. See for example [26]. The problem of computing the power of a test is much more complicated. One reason is that the hypothesis alternative to the one tested is usually composite (see subsection 7). In such a case, while the adopted level of significance is just one number α , the power of the test is not. If the composite alternative hypothesis splits itself into simple hypotheses, say $H_1, H_2, \dots, H_n, \dots$, then for each of them the power of a given test is likely to have a different value, say $\beta(H_1|\alpha)$, $\beta(H_2|\alpha), \dots, \beta(H_n|\alpha), \dots$. In fact, in this case it is appropriate to speak of a 'power function' of a test, rather than simply of its power. Furthermore, of course, the power function must depend upon the chosen level of significance, and there is a conflict: if one decreases α , then the power of the adopted test decreases also.

This particular circumstance is important in designing experiments. Not infrequently it happens that, with the contemplated number of observations and with the originally chosen rather low level of significance, i.e. $\alpha = 0.01$, one finds that the chance β of detecting 'errors' in the hypothesis tested judged large is rather small, say $\beta = 0.02$. Obviously, an experiment designed in this way is not worth performing. Unfortunately, this particular point escaped the attention of a large number of authors of statistical texts. Yes, the concept of power is occasionally mentioned, but its treatment is somewhat 'platonic'. As a result, important experiments are often performed with the possibility of very frequent most regrettable errors. An interesting paper on this subject has been recently published by Traxler [27].

In order to meet the situation in which the original design of an experiment yields $\alpha = 0.01$ and $\beta = 0.02$, there are several possible remedies, applicable singly or in combinations: (i) to alter the design of the experiment, (ii) to try to find a more powerful test, (iii) to increase the level of significance α and (iv) to increase the number of observations.

Incidentally, the customary parlance with reference to testing hypotheses involves two phrases which it is convenient to know. One phrase is that (some effect) is 'highly significant'; the other phrase is that the effect

in question is 'significant' or (perhaps) that it is 'significant at 5%'. The first phrase means that the hypothesis of no effect under discussion is rejected with the level of significance $\alpha = 0.01$. The meaning of the second is rejection with $\alpha = 0.05$, but not with $\alpha = 0.01$.

10. *Practical Applications of Theory of Testing Statistical Hypotheses*

With reference to the old dictum, 'the proof of the pudding is in the eating', the purpose of the present subsection is to visualise the working of the frequentist theory of testing hypotheses in a variety of studies of nature.

As emphasized above, the theory was born and constructed with the view of diminishing *the relative frequency of errors*, particularly of '*important*' errors. Thus, leaving aside the question of an error in testing some particular hypothesis, we have to contemplate a long sequence of situations, say $\{S_i\} = (S_1, S_2, \dots, S_n, \dots)$ in which tests of some hypotheses will be performed. This sequence, which we may label 'human experience', will be very heterogeneous. Some situations will refer to problems of astronomy [48], others to highway traffic, still others to radiation biology [49], some to problems of big cities and slums or to weather modification, etc. etc. However, there will be some elements common to all the situations of the sequence.

The elements common to all the situations typified by situation S_i will be: (1) a hypothesis H_i to be tested against an alternative \bar{H}_i , and (2) a subjective appraisal of the relative importance of the two kinds of error, leading to the adoption of an acceptably low level of significance α_i combined with an acceptable (hopefully 'optimal') power function. Let $\beta(H_i|\alpha_i)$ denote the value of this function corresponding to some specified simple alternative to H_i that may be judged important.

Eventually, then, with each situation S_i there will be connected a pair of numbers, α_i and $\beta(H_i|\alpha_i)$. The question is: what can one expect from the use of the theory of testing statistical hypotheses in the above heterogeneous sequence of situations summarizing human experience in 'pluralistic' studies of Nature? The answer is:

The relative frequency of first kind errors will be close to the arithmetic mean of numbers $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$ adopted by particular research workers as 'acceptably low' probabilities

of the more important errors to avoid. Also, the relative frequency of detecting the falsehood of the hypotheses tested, when false, and the contemplated simple alternatives happen to be true, will differ but little from the average of $\beta(H_1|\alpha_1)$, $\beta(H_2|\alpha_2), \dots, \beta(H_n|\alpha_n), \dots$.

This answer is a simple consequence of a theorem known as the central limit theorem of probability theory. An incredulous reader having access to a digital computer may wish to verify it empirically. The suggestion is to dream up a few hundred situations S_i , each with some statistical hypothesis H_i to be tested against an alternative \bar{H}_i . In order to mimic the 'human experience', the consecutive pairs (H_1, \bar{H}_1) , $(H_2, \bar{H}_2), \dots$, are likely to be all different but, for illustrative purposes, this is not necessary. In fact, all these pairs are not subject to any restriction.

The next step would be to decide on the hypothetical sequence of 'true states of nature', namely on cases where the hypothesis tested H_i will be true and where it will be false and how importantly false. There should be substantial numbers of cases of each kind, say at least 100. Consultations with a competent statistician will then determine the statistical test to be used in each case, the acceptably low significance level α_i and the value of the power $\beta(H_i|\alpha_i)$. Then help of a programmer will be needed to prepare the input for the high speed computer, etc.

All the above is emphasized at some length for a particular reason. This is that, at a variety of conferences with 'substantive scholars' (biologists, meteorologists, etc.), accompanied by their cooperating 'applied statisticians', I frequently hear a particular regrettable remark. This is to the effect that the frequency interpretation of either the level of significance α or of power β is only possible when one deals many times WITH THE SAME HYPOTHESIS H , TESTED AGAINST THE SAME ALTERNATIVE. Assertions of this kind, frequently made in terms of 'repeated sampling from the same population', reflect the lack of familiarity with the central limit theorem.

Is the above answer to the question of what to expect from the theory of testing hypotheses satisfactory? This is a subjective matter. On my own part, when faced by a hypothesis to be tested against an interesting alternative, I enjoy struggling for a criterion having just as high a power as possible. Also, I am uncertain about the possibilities of alternative ways of treating the many 'pluralistic' studies of Nature.

V. RANDOMIZATION OF EXPERIMENTS

11. *Basic Ideas*

Theory of frequentist statistics is so closely connected with that of experimentation with variable material that an article on the former must include at least some remarks on the latter. Here, a very important concept is 'randomization', introduced, emphasized and popularized by R. A. Fisher and his school, beginning with the mid 1920's [28, 29, 30]. It is concerned with an important pitfall in designing and conducting experiments.

As nicely described by Cochran [31], the experimentalists have a rather general trait: an emotional attachment to one or more subjects of their study. In an experiment with 'variable material' there are entities described as 'units of experimentation'. For example, in testing chemicals for carcinogenicity, mentioned earlier, the 'units of experimentation' are mice. In order to test a chemical A there must be some units of experimentation having no contact with A (these will be 'control' mice) and some others ('experimental' mice) exposed to A . Ordinarily, a statistically unsophisticated experimentalist will arbitrarily select several mice to be controls and some others to be experimentals. Depending upon his personal attitude (e.g., perhaps being attracted by the flavor of the food additive A , etc.) in selecting a sample of mice to serve as 'experimentals', the experimentalist will have a subconscious tendency to arrange that A is not unjustly diagnosed as 'carcinogenic'. To make sure, the experimentalist is likely to select his experimental mice out of those that look healthy and strong. Clearly, the experimentalist performing in this way is in danger of self-deception and in danger of deceiving others. The danger of such deception will be increased if the experimentalist happens to have some financial interest in marketing the additive A .

Fisher's important idea was that, in order to avoid errors in judging what is generally called the experimental 'treatments', the subdivision of the available 'units' to serve as experimentals and the controls must be made at random, not through an arbitrary choice of the experimentalist. This is what is called 'randomization' of an experiment.

The further idea is that the personnel involved in a randomized experiment should not be informed of which units are 'experimental' and which are 'controls'. Experiments so conducted are occasionally called

'double blind' experiments. Unfortunately, in many practical situations, the 'double blindness' of an experiment is difficult to achieve.

As of now, Fisher's idea that an experiment with variable material cannot be reliable unless it is randomized, has been accepted in many domains of science and technology. Generally, the acceptance came against strong opposition and after a substantial struggle (e.g., 'Oh, get this Fisher out of my hair! I know about my material and about my experiments all I need to know!')

One of the domains in which the struggle for randomization continues is weather modification through so-called cloud seeding. As might be expected, the principal opponents are the commercial enterprises specializing in cloud seeding [32].

The statistical hypotheses that come under test in connection with randomized experiments are of a special type. This type can be exemplified by the question: are the distributions of the observable variables corresponding to experimental and to control units identical or not? Alternatively, the same question is often worded as follows: how frequently can *the known scheme of randomization* produce differences between the experimental and the control units as great or greater than those observed? (Naturally, the term 'differences' requires specification, and it is here the problem of most powerful tests comes to the fore.)

12. *An Embarrassing Incident*

In order to emphasize the importance of Fisher's idea of randomization, it seems appropriate to mention a somewhat anecdotal case in our studies of cloud seeding experiments intended to verify the claims that the seeding (by a particular method and in specified conditions) tends to increase the average rainfall. In this particular case, with the experiment called the Whitetop, the experimental unit was a day satisfying certain weather conditions and cloud seeding was performed from three aircraft, supposed to fly back and forth over indicated points on the ground. The hope was to increase the rainfall over a specified circular 'target', at least over some points of this target.

Unfortunately, contrary to these hopes, the average rainfall on days with seeding was 'significantly' lower than that on control days [33]. Furthermore, it was found that this decrease in rain, ascribable to

seeding, occurred not only in the intended target, but over a huge area surrounding it, an area greater than the whole area of the United Kingdom. The estimated average decrease was about 20% of non-seeded days rainfall.

The descriptions of the widely publicized experiment included the statement that it was 'properly randomized in the statistical sense'. Therefore, initially, I had no doubts about the findings, but there was an intriguing question about the mechanisms in the atmosphere that could have produced the unexpected effect. A personal communication from James Hughes, a cloud physicist in the U.S. Office of Naval Research, suggested a possibility. This was connected with the change in cloudiness and temperature between the periods before the beginning of seeding (10–11 o'clock before noon) and after. When trying to verify this hypothesis we ran into most unexpected facts relating to precipitation during the ten hours BEFORE the commencement of seeding.

It appeared [34] that in three vast areas the average seed–no seed rainfall difference was 'approximately significant', 'significant' or 'highly significant', respectively. Furthermore, all these differences were negative: less rain on seeded than on not seeded days. Still more surprising, these differences occurred in areas 90 to 180 miles away from the target, either directly or partly upwind. The seeded day average precipitation in these three areas was about $\frac{2}{5}$, $\frac{1}{3}$ and $\frac{1}{5}$, respectively, of that without seeding! Since these differences could not have been caused by seeding and could hardly be ascribed to chance in the process of faultless randomization, the inescapable conclusion is that there must have been some flaw in the strict randomization.

Some reading of climatological literature brought to light the fact that, depending upon winds, not infrequently the three areas in question must have included the Ozark Plateau, which some authors described as the 'breeding ground' of severe storms. Could it be that the early morning weather reports indicating particularly stormy weather in upwind areas caused the abandonment of seeding, even though the randomized decision was to seed? Who knows? In any case, no reliable conclusions about the effects of seeding in this experiment appear possible.

When thinking of designs of experiments, Fisher visualized many difficulties and invented ways of circumventing them. The experience with the Whitetop trial, which lasted five years and must have cost the

U.S. Government some millions of dollars, suggests an idea which Fisher overlooked. This is that an important experiment should be so designed that the personnel involved in it do not find it too uncomfortable to adhere to strict randomization. Flying a small plane in stormy weather, particularly flying through a thunderstorm cloud, must be unpleasant and may well be dangerous. Thus, the method of cloud seeding used in the Soviet Union, by firing rockets or artillery shells into selected points within the clouds, is likely to be preferable to the use of aircraft, unless they can fly high above the cloud top where the turbulence is less pronounced. Regretfully, in spite of an excellent school of probability in the Soviet Union, their cloud seeding experiments known to me are not randomized.

In this connection, it is appropriate to mention a recent article [35] on the subject of 'Who Needs Randomization?' The answer is: the randomization is necessary to the 'consumer' of the results of an experiment. In particular, if a costly experiment is financed by a governmental institution, then, along with the customary auditing of the expenditures, the institution concerned should insist on randomization, and should 'audit' the process of randomization. Better still, it should 'monitor' the randomization. Mere assertions to the effect that 'the experiment was properly randomized in the statistical sense' are not enough.

VI. SKETCH OF THE PROBLEM OF STATISTICAL ESTIMATION

The following examples are intended to illustrate two different kinds of statistical problem of estimation, as they occur in a variety of studies of nature.

13. *Example of an Isolated Problem*

For some reason, the government of a country needs data on all the farms in its territory, a very large number N of them. The data needed relates to some date in the not very distant past, say 1 January 1976. Among other things, the government needs a number θ , perhaps some average. The errorless value of θ can be obtained through a complete survey of all the N farms and a detailed study of each of them. Because of the time needed,

a complete survey of all the N farms is impractical. Instead, a sample survey of a much smaller number n of farms is contemplated, using one or another of the developed methods of random sampling [36, 37]. Let X_1, X_2, \dots, X_n denote the relevant data for the n farms of the sample, and let a single letter X denote their totality. Before the sample is taken the values of X are not determined and the chance mechanism of random sampling, jointly with the characteristics of the population of farms, determine the variability of X or, as we call it, 'the distribution of X '. For this reason, the variables X are called 'observable random variables'.

The mathematical-statistical problem of estimation of θ consists in devising methods whereby the observable random variables X could be used to obtain reliable information on the value of θ .

Note that, while the problem involves a known chance mechanism governing the variability of X , no such chance mechanism affects θ . θ is just a fixed, but an unknown, number.

14. *Example of Connected Problems Treated Routinely*

The daily routine of a clinic with many customers includes the analysis of the blood sugar content. Denote by θ_i the true blood sugar content of the i th patient. All analyses are subject to error and the clinic performs a number n of parallel analyses for each patient. Let $X(i) = [X_{i1}, X_{i2}, \dots, X_{in}]$ denote the results of such analyses performed for the i th patient. We shall proceed on the assumption that the chance mechanism of variability of the $X(i)$ is sufficiently established by the so-called 'theory of errors'.

The theoretical-statistical problem of estimation relating to the i th patient in the clinic consists in developing a method of using the observable random variables $X(i)$ in order to obtain reliable information regarding his true blood sugar content θ_i .

At first sight, the problem of estimation in the two examples is identical. In both cases one has to do with some unknown number θ or θ_i and several observable random variables X or $X(i)$, the chance mechanism governing their variability assumed known. However, there is a difference. In Example 1 we deal with just one (as we call it) 'parameter' θ to be estimated. Contrary to this, Example 2 involves routine analyses of blood sugar content for patients, the number of which, over a year, may

well run into thousands. Depending upon circumstances, among these patients certain categories can be established, each with a rather fixed condition of life, perhaps coal miners or office workers, which influence the frequencies of the blood sugar content θ_i having small or large values.

In other words, Example 2 differs from Example 1 by the fact that the consecutive values of the blood sugar content θ_i can be considered as particular values of a certain random variable which we shall denote by Θ . Naturally, if the distribution of Θ were known, this information could be used in order to improve the precision of estimating each θ_i . It is here that the famous formula of Bayes comes to the fore with the distribution of Θ playing the role of the so-called 'prior'. Unfortunately, however, while it is appropriate to consider the successive values of the θ_i as particular values of a random variable Θ , the distribution of this variable is not known and there are obvious difficulties to establish it (see point (i) in Section 2). The difficulty is that the analyses performed for the consecutive patients in the clinic do not give us the values of their true blood sugar content θ_i , but only the determinations $X(i)$, and the transition from the distribution of the latter to that of the former constitutes quite a problem.

15. *Empirical Bayes Theory*

Naturally, if the problem of Example 1 is solved, then this solution can be applied to problems illustrated in Example 2. But it is also natural to try to do better. A brilliant idea as to how it can be done, initiating a novel chapter of frequentist mathematical statistics, is due to Herbert Robbins [38]. The new theory of Robbins is known under the title Empirical Bayes Theory. See also [39].

16. *Point Estimation*

Problems of Example 1 within the frequentist theory of statistics fall under two headings: 'point estimation' and 'confidence intervals'. The first of these has a long history, going back to Laplace and particularly to Gauss, and is at the base of the theory of least squares. Briefly and roughly, it consists in determining a function, say $\theta^*(X)$, of the observable variables (now termed 'estimator') the values of which are frequently 'close' to the estimated θ . In many cases, the theory also provides

information on the frequency distribution of the 'error' $\theta^*(X) - \theta$. In frequent cases (but not always) an excellent 'point' estimator $\theta^*(X)$ is found by maximizing the so-called 'likelihood function', a concept due to R. A. Fisher.

The theory of confidence intervals is sketched in the next section.

VII. SKETCH OF THE THEORY OF CONFIDENCE INTERVALS

17. *Basic Concepts*

With reference to the preceding section we denote by θ a fixed but unknown number, to be estimated using certain observable random variables $X = (X_1, X_2, \dots, X_n)$, the distribution of which depends on θ . The datum of the problem includes the set of values that θ can possibly have. Usually, but not always, it is given that θ can be any number between some given limits a and b , such as $a = 0$ and $b = 1$ or $a = 1$ and $b = +\infty$ etc.

The other datum of the problem is the distribution of the observables X depending upon the value of the parameter θ .

The problem of confidence intervals consists in determining two functions of the observables, say $Y_1(X)$ and $Y_2(X)$, satisfying the inequalities $a \leq Y_1(X) < Y_2(X) \leq b$ to be used in the following manner: Whenever the observable variables X assume some values $x = (x_1, x_2, \dots, x_n)$, we shall calculate the corresponding values of Y_1 and Y_2 , say $Y_1(x) < Y_2(x)$, and then assert (or act on the assumption) that

$$(1) \quad Y_1(x) \leq \theta \leq Y_2(x).$$

The two functions $Y_1(X)$ and $Y_2(X)$ are called the lower and the upper confidence limits (or 'bounds') for θ and the interval between them, say $I(X) = [Y_1(X), Y_2(X)]$ the confidence interval.

Being functions of the random variable X , the two confidence bounds and the confidence interval $I(X)$ will be random variables also. (Here certain conditions of 'measurability' must be imposed, but they are customarily satisfied and are too technical for the present article.) In order to be useful as tools of inductive behavior, the confidence bounds, and the interval $I(X)$ between them, must possess certain well defined frequency properties.

One important property is that the assertions about the unknown number θ typified by the double inequality (1) be FREQUENTLY correct, and this irrespective of the value that θ may possess. Accordingly, we select a positive number $\alpha < 1$, acceptably close to unity, and require that the probability of the two confidence bounds 'bracketing' the true value of θ be identically equal to α whatever the value of θ may be. This requirement is expressed by the formula:

$$(2) \quad P\{Y_1(X) \leq \theta \leq Y_2(X) | \theta\} \equiv \alpha.$$

The number α used is called the 'confidence coefficient'. The frequently used values of α are 0.90, 0.95 and 0.99, for which convenient numerical tables are available.

The theory published since late 1930's [40, 41, 42] indicates the possibility of satisfying the identity (2) in many different ways. This circumstance poses the question of which of the many (usually, an infinity) of different pairs of bounds $[Y_1(X), Y_2(X)]$ to use. In other words, just as in the problem of statistical tests, there is the question of optimality of the confidence intervals, all corresponding to the same chosen confidence coefficient α . As might be expected, there have been conceived quite a few concepts of optimality, depending upon the nature of the applied problem and, frequently, on the difficulty of the relevant mathematical problem of reaching the desired 'optimum'.

One intuitively easy definition of optimality is that the length of the confidence interval, say

$$(3) \quad L(X|\alpha) = Y_2(X) - Y_1(X)$$

be, in a sense, just as small as possible without the infringement of the basic requirement (2). However, even with this apparently simplest condition of optimality there are delicate conceptual points. They are connected with the fact that the difference in the right side of the formula (3) is a random variable of which it is only known that all its possible values are positive. This being the case, what does it really mean to require that this difference be 'as small as possible'? One possibility is to require that the AVERAGE value of $L(X|\theta)$ be a minimum. In mathematical terms this would mean the requirement that the 'expectation' of $L(X|\alpha)$ be a minimum. However, the requirement that the expectation of $L(X|\alpha)$ be a minimum is not the only way of defining the optimum.

The reader will realize that the above descriptions of the problem of confidence intervals involves oversimplifications, necessary to make the presentation 'uncluttered' by technicalities and yet emphasizing the basic concepts. Further below a simple example is likely to be helpful. Here, it is appropriate to mention that, in some cases, an exact satisfaction of the basic identity (2) is impossible (i.e. without introducing certain artificialities). This occurs when the observable variables X are so-called 'discrete'. In cases of this kind, rather than require the exact equality to α in formula (2) one can require 'at least equal' or 'approximately equal', etc. Another important technical difficulty arrives when the distribution of the observables X depends on not just one parameter θ that one wants to estimate, but on several of them, the values of which are of no particular interest. On the initiative of the late Harold Hotelling, such additional, not immediately interesting, parameters are called 'nuisance parameters': they interfere with the problem of estimating θ which is of prime interest.

18. *Anticipated Misunderstandings*

Before proceeding to the construction of confidence intervals, we must discuss the meaning of formulas (1) and (2) and anticipate certain misunderstandings.

The important point is the distinction between the symbols X and x . The first denotes a set of the observable random variables $X = (X_1, X_2, \dots, X_n)$ the variation of which is governed by a chance mechanism which, in some specified way, depends upon the unknown number θ . The second symbol $x = (x_1, x_2, \dots, x_n)$ denotes some n numbers which, in some particular case, may have been assumed by the random variables, so that $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$. When this occurs, our inductive behavior would require us to assert (1). However, the basic formula (2) is written not in terms of the *observed* x but in terms of the *observable* X . Barring some blunders in calculations, etc. this formula is true whatever may be the value of the unknown θ . However, if one substitutes in (2) the *observed* x in the place of the *observable* X , the result would be absurd. In fact, the numerical results of the substitution may well be

$$(4) \quad P\{Y_1(x) \leq \theta \leq Y_2(x) | \theta\} = P\{1 \leq 5 \leq 3 | 5\} = 0.95,$$

or alternatively,

$$(5) \quad P\{1 \leq 2 \leq 3 | 2\} = 0.95.$$

It is essential to be clear that both (4) and (5) are wrong. The probability in the left hand side of (4) has the value zero (and thus not 0.95), and that in the left hand side of (5) is unity, neither of any interest.

The meaning of the identity (2) is that a systematic use of confidence bounds $Y_1(X)$ and $Y_2(X)$, whether in estimating one particular unknown θ , or all different, without any restriction, the RELATIVE FREQUENCY of correct assertions will be close to the selected $\alpha = 0.95$.

Here is another kind of misunderstanding against which the reader must be warned. Consider two practicing statisticians who happen to treat the same problem of estimating an unknown θ at the same confidence coefficient α , and are faced with some already observed $x = (x_1, x_2, \dots, x_n)$. However, the two statisticians differ in their concepts of optimality and the confidence bounds they use are different, say $[Y_1^*(X), Y_2^*(X)]$ and $[Y_1^{**}(X), Y_2^{**}(X)]$, respectively.

Having the data x provided by the observations, the assertions about θ of the two statisticians may well be:

$$(6) \quad 1 \leq \theta \leq 3 \quad \text{and} \quad 4 \leq \theta \leq 5,$$

respectively. Another possibility is

$$(7) \quad 1 \leq \theta \leq 3 \quad \text{and} \quad 2 \leq \theta \leq 4.$$

The pair of assertions (6) is contradictory and, if the two statisticians are employed as consultants, perhaps in the same government agency, the practical problem of whom to believe may be quite acute. The proper solution is for the 'consumer' to understand a little the long run frequency properties of the two pairs of confidence bounds used and then to make an informed choice.

As to the difference between two assertions exemplified in (7) I have seen occasions in which such differences did occur and where the practical conclusion was reached that the unknown θ must be included in the common part of the two intervals, namely $2 \leq \theta \leq 3$. At the time when this conclusion was reached, there was no theoretical basis supporting it and I am not sure whether it exists now. However, it may be interesting to investigate the frequency properties of the procedure involving the

calculation of two different specified confidence intervals and, in cases when they overlap, to use their common part as some kind of 'summary' confidence interval.

19. *Construction of Confidence Intervals: Regions of Acceptance*

The purpose of this subsection is to provide the reader with an intuitive feeling about the general problem of constructing confidence intervals. In order to avoid inessential technicalities connected with the consideration of multidimensional spaces, we shall assume that the number n of observable variables is $n = 1$. Alternatively, we may consider the case where it is decided to base the estimation of θ on some selected function of $n > 1$ observables, such, for example, as their arithmetic mean.

The horizontal axis in Figure 1 is used to measure the possible values x of X , the single observable random variable. For the sake of simplicity, we shall assume that all possible values of X are those in the interval from 0 to some number M . Thus, the segment from zero to M of the horizontal axis represents what is called the 'sample space' of X . The vertical axis is reserved for all possible values of the parameter θ , all of them between the indicated limits $a < b$.

The three points marked $\varphi_1, \varphi_2, \varphi_3$ on the vertical axis symbolize three particular values that θ may possess. The horizontal lines through these points are replicas of the sample space of X . The fancy curve above the lowest of these lines is meant to represent or symbolize the probability density of X as determined by $\theta = \varphi_1$. A somewhat different curve is drawn above the line corresponding to φ_3 . A glance at the two curves indicates that, if $\theta = \varphi_3$, then X will assume larger values somewhat more frequently than with $\theta = \varphi_1$. Whatever the case may be and whatever be the true value of θ , on each of the horizontal lines like those corresponding to φ_1, φ_2 and φ_3 there will be an infinity of intervals, say $A(\varphi)$ such that the probability of X falling within $A(\varphi)$, as determined by $\theta = \varphi$, will be exactly equal to the chosen α . Intervals having this property are called 'regions of acceptance'. $A(\varphi_1)$ and $A(\varphi_3)$ are indicated in Figure 1.

Now visualize that for every possible value φ of θ we selected arbitrarily a region of acceptance $A(\varphi)$. Next, visualize a line, say $\varphi = Y_1(x)$, connecting the right ends of all these regions and another line $\varphi = Y_2(x)$ connecting their left ends. Depending on how the particular $A(\varphi)$'s are

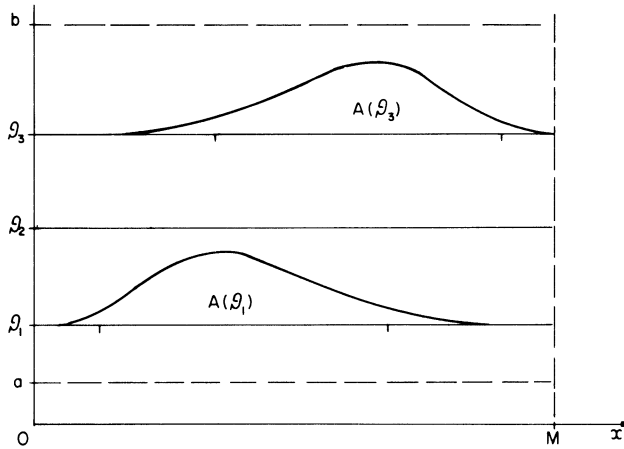


Fig. 1. Illustration of the concept of 'regions of acceptance' $A(\varphi)$. Quantity measured on the horizontal axis is x = possible value of the observable r.v. X . Quantity measured on the vertical axis is φ = possible value of the estimated parameter θ .

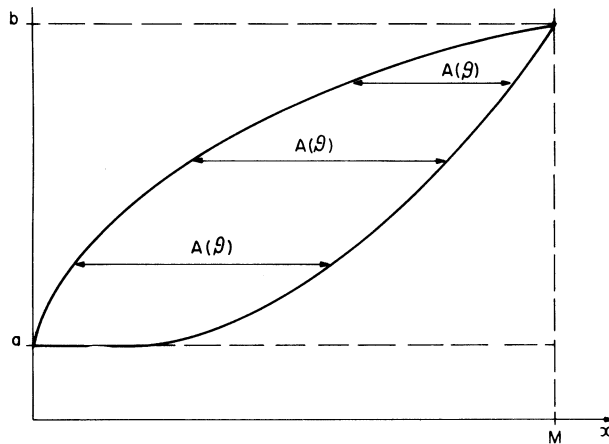


Fig. 2. Illustration of the 'confidence belt': the curve on the left passes through the left boundaries of acceptance regions $A(\varphi)$ corresponding to all the possible values φ of the estimated parameter θ . Similarly, the curve on the right passes through the right boundaries of the same $A(\varphi)$. Equations of the two curves can be solved either with respect to x or with respect to φ . The latter solutions $Y_1(x)$ and $Y_2(x)$ give the confidence interval corresponding to $X = x$.

selected, the two lines determined by $Y_1(x)$ and $Y_2(x)$ may be very fancy and discontinuous. However, since the particular $A(\varphi)$ are subjected to a single condition that the probability determined by φ of X falling into $A(\varphi)$ is equal to preassigned α , it is obvious that, through a little pushing of the $A(\varphi)$ one can achieve a degree of regularity of the two curves.

Now, assume that this is done and have a glance at Figure 2 exhibiting the two fairly regular curves selected by us. Remembering that the vertical axis in Figure 1 is allocated to possible values φ of the estimated θ , the equations of the two lines may be written as

$$(8) \quad \varphi = Y_1(x) \quad \text{and} \quad \varphi = Y_2(x)$$

The construction of a confidence interval for estimating θ at the chosen confidence coefficient is now complete. In fact, substitute in (8) the random variable X for its value x and examine the probability that $Y_1(X)$ and $Y_2(X)$ as defined through (8) will bracket the unknown true value of θ . Obviously, for this to happen it is both necessary and sufficient that the observable random variable X falls within the region of acceptance $A(\theta)$. But these regions were purposefully selected to ensure that the probability in question be equal to α ! Thus, the probability that the two bounds $Y_1(X)$ and $Y_2(X)$ will bracket the unknown θ is also equal to α .

20. *An Example*

A book of instructions about how to drive an automobile is a very useful source of information. However, those who learned to drive are sure to agree that the real feel of what is involved is acquired only after a few exercises on the road. The following example is offered in this spirit. In order to be non-technical, the problem considered is 'bookish', without contact with any real study of nature.

It is given that the n observable random variables X_i are mutually independent and are uniformly distributed within an interval $(0, \theta)$ of unknown length θ . It is this length θ that is the parameter to be estimated by confidence intervals corresponding to a preassigned confidence coefficient α , say $\alpha = 0.90$. It is also given that θ may have any value, say $50 \leq \theta \leq 600$. (These numerical limits are included in order to ensure an appropriate degree of approximation, suggested below, and for convenience in certain graphs.)

Leaving aside the dealing with all the n variables, let us visualize two possibilities of summarizing them by just one function. One suggestion is that the arithmetic mean, say \bar{X} , of the n observable variables might be a good 'summary' variable to estimate θ . The other suggestion is to use for the same purpose the greatest of the n observations, say X^* .

With a moderate value of n , say $n = 10$, the distribution of \bar{X} will be approximately 'normal', centered at $\theta/2$, with a variance equal to $\theta^2/12n = \theta^2/120$. This circumstance suggests the possibility of adopting as the region of acceptance, say $A_1(\bar{X})$, the symmetric interval about $\theta/2$, of an appropriate length so chosen that the probability of \bar{X} falling in it be equal to $\alpha = 0.90$. In other words, our intuitively selected region of acceptance A_1 will extend from $\theta/2 - k$ to $\theta/2 + k$, where k is to be adjusted to the chosen value of the confidence coefficient α . Using the normal approximation to the distribution of \bar{X} the value of k is easily found to be $k = (0.15)\theta$. Thus, whatever θ may be, the probability that \bar{X} will fall within the interval

$$(9) \quad \theta/2 - (0.15)\theta \leq \bar{X} \leq \theta/2 + (0.15)\theta$$

is approximately equal to $\alpha = 0.90$.

The symbols $Y_1(X)$ and $Y_2(X)$ introduced in the preceding subsection denote, respectively, the smallest and the largest values of θ for which X falls within the region of acceptance corresponding to that θ . The right part of the double inequality (9) indicates that the least value of θ for which a given \bar{X} does not exceed the right boundary of the acceptance region is equal to $(1.548)\bar{X}$. This is, then, our $Y_1(X)$ or, as we shall denote it now, $Y_1(\bar{X})$. Similarly, the greatest value of θ for which an observed \bar{X} will be at least equal to the left boundary of the corresponding acceptance region is, say, $Y_2(\bar{X}) = (2.857)\bar{X}$. It follows that, in the present case, the two curves sketched in Figure 2 as connecting the boundaries of acceptance regions, are simply two straight lines passing through the origin with slopes approximately equal to 1.54 and to 2.86, respectively. They are exhibited in Figure 3.

The use of this Figure is as follows. One observes the $n = 10$ variables as defined at the outset, one calculates their arithmetic mean, say \bar{x} , and one draws a vertical line through the point \bar{x} on the horizontal axis. The intersections of this vertical with the two sloping lines determine the confidence interval $[Y_1(\bar{x}), Y_2(\bar{x})]$. This is illustrated in Figure 3.

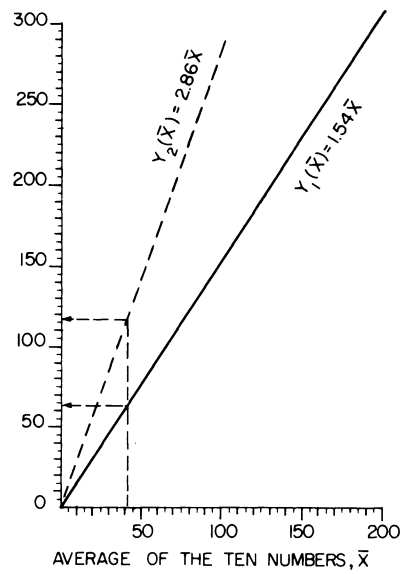


Fig. 3. Diagram facilitating the determination of the confidence intervals based on \bar{X} .

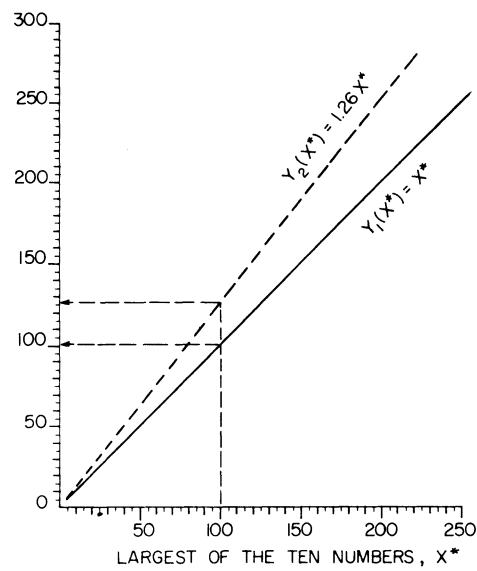


Fig. 4. Diagram facilitating the determination of the confidence interval based on X^* .

Now, consider the second suggestion, namely that the ten observations on X be 'summarized' by the greatest of them, denoted by X^* . Here, then, the probability that X^* will not exceed a preassigned number $t < \theta$ coincides with the probability that each and every one of the n observations will not exceed t . Obviously this probability is exactly equal to $(t/\theta)^n$. This is, then, the exact distribution function of the random variable X^* .

As to a plausibly good region of acceptance, say $A_2(\theta)$, based on X^* , one might take into consideration the fact that X^* cannot possibly exceed θ . Thus, what about defining $A_2(\theta)$ as an interval ending at θ and beginning with a value t such that $(t/\theta)^n = 1 - \alpha$? Simple calculations show that with $n = 10$ and $\alpha = 0.90$, the requisite value of t equals θ divided by the tenth root of 10. Calculations similar to those leading to the confidence interval based on \bar{X} yield, say

$$(10) \quad Y_1(X^*) = X^* \quad \text{and} \quad Y_2(X^*) = (1.26)X^*.$$

Confidence intervals determined by (10) can be read directly from Figure 4.

Now we come to the important question as to what might be the long run results of using the two confidence intervals $[Y_1(\bar{X}), Y_2(\bar{X})]$ and $[Y_1(X^*), Y_2(X^*)]$ corresponding to the same $\alpha = 0.90$. The anticipated answer is that both intervals will bracket the true value of θ with about the same relative frequency of 90%. But will they? The other interesting question is which of the two intervals will offer the better 'precision' of estimation?

In order to answer these questions empirically a total of 100 experiments were simulated by the Monte Carlo technique and Figures 5 and 6 exhibit the results.

Each of the 100 experiments consisted in adopting some value, of $\theta = 50, 100, 150$ etc. Next, for each of these values the high speed computer simulated 20 experiments, each of $n = 10$ observations, on a random variable uniformly distributed between zero and θ . Then the arithmetic mean of these observations was used to calculate the confidence interval from formula (9). The vertical lines in Figure 5 exhibit these intervals which makes it possible to count how many times they cross the horizontal line that marks the assumed true value of θ . Similarly, Figure 6 exhibits the confidence intervals based on the greatest of the ten observations in each sample.

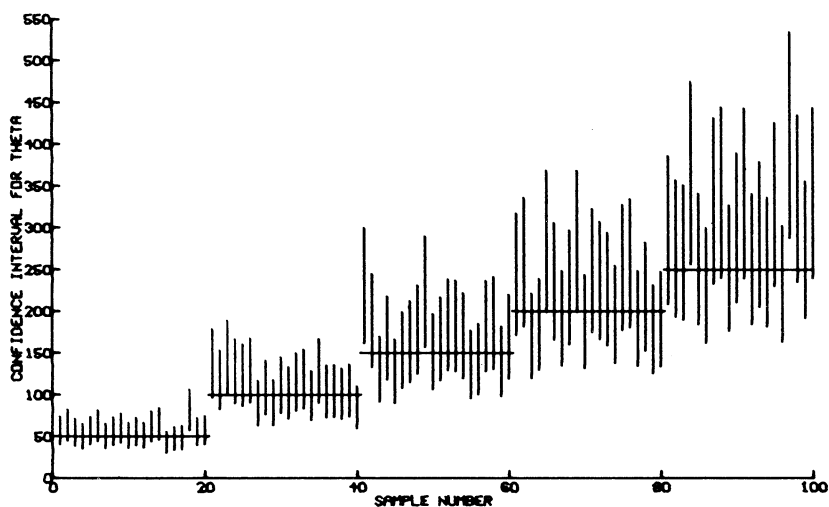


Fig. 5. High speed computer output simulating 100 experiments, each yielding a confidence interval for θ based on \bar{X} . Solid vertical lines give the confidence intervals obtained for each of the 100 samples.

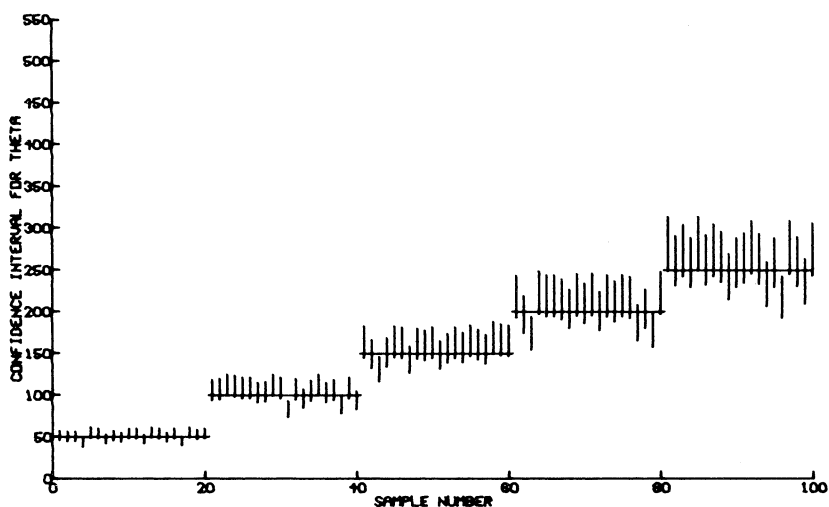


Fig. 6. High speed computer output simulating 100 experiments (same as in Figure 5), each yielding a confidence interval for θ based on X^* . Solid vertical lines give the confidence intervals obtained for each of the 100 samples.

A glance at the two Figures indicates an affirmative answer to the question about the frequency of each confidence interval bracketing the true value of θ : the actual frequency is pretty close to 90%. But what about the ‘precision’ of estimating θ ? Another glance at the two Figures should give an emphatic answer that the precision in the two cases is not the same. Specifically, the precision of estimating θ by using X^* is much better than that using the mean \bar{X} . This is the justification of the theoretical efforts (a) to define appropriately the ‘optimality’ of the confidence intervals, and (b) to develop the methodology of reaching or, at least, approaching the optimum. Depending upon the nature of the experiment and/or the distributions of the observable variables, these mathematical problems can be very difficult.

21. *Concluding Remark*

The reader will notice that the properties of the confidence intervals illustrated in Figures 5 and 6 are entirely independent of whether the true values of θ are just constants as in subsection 13, or are particular values of some random variable Θ as in subsection 14.

VIII. A GLANCE AT THE HISTORY OF ESTIMATION BY CONFIDENCE INTERVALS OR REGIONS

The first authorship of a scholarly idea of some delicacy is extremely difficult to establish. The first publications involving the basic ideas of confidence intervals known to me [43, 44] appeared in 1929 and 1931, both concerned with particular problems, as reflected in their titles: ‘Applications of the Theory of Errors to the Interpretation of Trends’, and ‘The Generalization of Student’s Ratio’, respectively. However, the basic idea that randomness and the calculation of the probabilities refer to the estimators rather than to the estimated parameters is contained in these papers. The more important of these papers appears to be that of 1931 due to Hotelling. This paper is concerned with simultaneous estimation of not just one unknown parameter but several of them, namely of means of several possibly correlated normal variables.

For quite some time the results of Hotelling remained unnoticed, but later their importance became increasingly appreciated, particularly due

to the idea of Henry Scheffé [45] concerned with the ‘multiple comparison problem’. A more recent book on the subject [46] is due to Rupert Miller. Predominantly, this literature is concerned with random variables assumed normally distributed. Another outstanding characteristic of these papers is that, from the very start, the ideas involved were entirely free of Bayesianism: the estimated quantities were unknown constants, not subject to prior distribution.

My own involvement in the field was also due to the necessity of solving an applied problem, in which a student of mine, Waclaw Pytkowski was involved. However, contrary to Hotelling, I began as a quasi-Bayesian. My assumption was that the estimated parameter (just one!) is a particular value of a random variable having an unknown prior distribution. My efforts were directed towards an estimator interval with its probability of covering the true value of the estimated parameter being independent of the prior distribution [36].

The elements of the theory I developed were the subject of my lectures first in Warsaw and, in 1934, also at the University College, London. Pytkowski’s work [47] acknowledging my lectures on confidence intervals, was published in 1932. My first publication [36] mentioning and briefly discussing confidence intervals appeared in 1934. The first basic results appeared in 1937–38 [40, 41]. From one point of view my first results were less general than those of Hotelling: I was concerned with estimating just one parameter while Hotelling estimated several of them jointly. However, one particular aspect of my results is more general than that of Hotelling: my results were not tied to normal, or to any other distribution of the observable variables.

Here it is a pleasure to acknowledge help of another of my former students, namely Churchill Eisenhart. He attended my lectures at the University College, London, and witnessed my introducing a prior distribution of the estimated θ and then making efforts to produce an interval estimator, the properties of which would be independent of the prior. Once, Eisenhart’s comment was that the whole theory would look nicer if it were built from the start without any reference to Bayesianism and priors. This remark proved inspiring.

Statistical Laboratory, University of California, Berkeley

NOTE

* The present paper was prepared using the facilities provided by three grants: the U.S. Energy Research and Development Agency; the National Institutes of Health, research grant No. ESO1299-13; the Office of Naval Research, contract No. NOOO14-75-C-0159/NRO82-230. I am indebted to Mr. Keith Sharp for performing the Monte Carlo simulation experiment which produced Figures 5 and 6.

REFERENCES

- [1] de Finetti, B., *Probability, Induction and Statistics*, John Wiley & Sons, New York, 1972.
- [2] *A Selection of Early Statistical Papers of J. Neyman*, University of California Press, Berkeley, 1967.
- [3] Neyman, J., 'The Emergence of Mathematical Statistics', in *On the History of Statistics and Probability* (ed. by D. B. Owen), Marcel Dekker, New York, 1976.
- [4] Neyman, J., 'On the Use of Maximum Likelihood Estimators', *Bulletin of the International Statistical Institute* **38**, Part 1 (1961), 193–200.
- [5] Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. 1, John Wiley & Sons, New York, 3rd ed., 1968.
- [6] Harris, T. E., *The Theory of Branching Processes*, Springer-Verlag, Berlin, 1963.
- [7] Karlin, S., *A First Course in Stochastic Processes*, Academic Press, New York, 3rd ed., 1969.
- [8] Kolmogorov, A. N., *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Julius Springer, Berlin, 1933.
- [9] Doob, J. L., *Stochastic Processes*, John Wiley & Sons, New York, 1953.
- [10] Dynkin, E. B., *Markov Processes*, Springer-Verlag, Berlin, 1965.
- [11] Loève, M., *Probability Theory*, Van Nostrand, New York, 2nd ed., 1960.
- [12] von Mises, R., *Wahrscheinlichkeit Statistik und Wahrheit*, Julius Springer, Vienna, 1936. See also von Mises, R., *Probability Statistics and Truth* (trans. H. Geiringer), George Allen and Unwin Ltd., London, 1957.
- [13] Borel, E., *Elements de la Théorie des Probabilités*, Hermann, Paris, 3rd ed., 1924.
- [14] Borel, E., *Le Hasard*, Hermann, Paris, 1920.
- [15] Neyman, J. and Pearson, E. S., 'On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference', *Biometrika* **20-A**, Part 1 (1928), 175–240. (See also [17] 1–66.)
- [16] Neyman, J. and Pearson, E. S., 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', *Philosophical Transactions of the Royal Society of London, Series A* **231** (1933), 289–337. (See also [17] 140–185.)
- [17] *Joint Statistical Papers of J. Neyman and E. S. Pearson*, University of California Press, Berkeley, 1967.
- [18] Neyman, J., *Lectures and Conferences on Mathematical Statistics and Probability*, Graduate School of U.S. Department of Agriculture, Washington, 2nd ed., 1952.
- [19] Brownlee, K. A., *Statistical Theory and Methodology in Science and Engineering*, John Wiley & Sons, New York, 1960.
- [20] Lehmann, E. L., *Testing Statistical Hypotheses*, John Wiley & Sons, New York, 1959. Translated into Russian, Polish, and Japanese.

- [21] Schmetterer, L., *Einführung in die Mathematische Statistik*, Springer-Verlag, Vienna, 1966.
- [22] Berger, A. and Wald, A., 'On Distinct Hypotheses', *Annals of Math. Stat.* **20** (1949), 104–109.
- [23] Neyman, J. and Pearson, E. S., 'The Testing of Statistical Hypotheses in Relation of Probabilities *A Priori*', *Proc. Cambridge Philos. Soc.* **29** (1933), 492–510.
- [24] Wald, A., *Statistical Decision Functions*, John Wiley & Co., New York, 1950.
- [25] Davies, R. B., 'Beta-Optimal Test and an Application to the Summary Evaluation of Experiments', *J. of the Royal Statistical Society, Series B* **31** (1969), 524–538.
- [26] Neyman, J., 'Optimal Asymptotic Tests of Composite Statistical Hypotheses', *Probability and Statistics* (The Harald Cramér Volume) (ed. by U. Grenander), Almqvist and Wiksells, Uppsala, Sweden, 1959, pp. 213–234.
- [27] Traxler, R. H., 'Snag in the History of Factorial Experiments', in *On The History of Statistics* (ed. by D. B. Owen), Marcel Dekker, New York, 1976, pp. 281–295.
- [28] Fisher, R. A., *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, 5th ed., 1934.
- [29] Fisher, R. A., *The Design of Experiments*, Oliver and Boyd, Edinburgh, 1936.
- [30] Fisher, R. A. and Yates, F., *Statistical Tables for Biological, Agricultural and Medical Research*, Hafner, New York, 6th ed., 1963.
- [31] Cochran, W. G., 'The Vital Role of Randomization in Experiments and Surveys', in *The Heritage of Copernicus* (ed. by J. Neyman), Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1974, pp. 445–463.
- [32] Neyman, J., 'Experimentation With Weather Control', *J. of the Royal Statistical Society* **130** (1967), 285–326.
- [33] Neyman, J., Scott, E. L., and Smith, J. A., 'Areal Spread of the Effect of Cloud Seeding at the Whitetop Experiment', *Science* **163** (1969), 1445–1449.
- [34] Neyman, J., Lovasich, J. L., Scott, E. L. and Wells, M. A., 'Hypothetical Explanations of the Negative Apparent Effects of Cloud Seeding in the Whitetop Experiment', *Proc. U.S. Nat. Acad. Sci.* **68** (1971), 2643–2646.
- [35] Harville, D. A., 'Experimental Randomization: Who Needs It?', *The American Statistician* **29** (1975), 27–31.
- [36] Neyman, J., 'On the Two Different Aspects of the Representative Method', *J. Royal Stat. Soc.* **97** (1934), 558–625. (Spanish version of this paper appeared in *Estadística, J. Inter-American Stat. Inst.* **17** (1959), 587–651.)
- [37] Hansen, M. H. and Madow, W. G., 'Some Important Events in the Historical Development of Sample Surveys', in *On the History of Statistics and Probability* (ed. by D. B. Owen), Marcel Dekker, New York, 1976, pp. 73–102.
- [38] Robbins, H., 'An Empirical Bayes' Approach to Statistics', in *Proc. Third Berkeley Symp. Math. Stat. and Prob.*, Vol. 1, University of California Press, Berkeley, 1956, pp. 157–164.
- [39] Neyman, J., 'Two Breakthroughs in the Theory of Statistical Decision Making', *Rev. of the Intern. Stat. Inst.* **30** (1962), 11–27. (In Spanish in *Estadística Espanola* **18** (1963), 5–28; in Russian in *Matematika* **2** (1965), 113–132; in Bulgarian in *Phys. Math. Journ., Bulgarian Acad. Sci.* **10** (1967), 94–110.)
- [40] Neyman, J., 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability', *Philosophical Transactions of the Royal Society of London, Series A* **236** (1937), 333–380.

- [41] Neyman, J., 'L'estimation Statistique Traitée comme un Problème Classique de Probabilité', *Actualities Scientifiques et Industrielles* **739** (1938), 25–57. (Russian version of this paper appeared in *Uspehi Matematicheskikh Nauk* **10** (1944), 207–229.)
- [42] Neyman, J., 'Foundation of the General Theory of Statistical Estimation', *Actualities Scientifiques et Industrielles* **1146** (1951), 83–95.
- [43] Hotelling, H. and Working, H., 'Applications of the Theory of Errors to the Interpretation of Trends', *J. American Statistical Association* **24** (1929), 73–85.
- [44] Hotelling, H., 'The Generalization of Student's Ratio', *Annals of Math. Stat.* **2** (1931), 360–378.
- [45] Scheffé, H., 'A Method for Judging All Contrasts in the Analysis of Variance', *Biometrika* **40** (1953), 87–104.
- [46] Miller, R. G., *Simultaneous Statistical Inference*, McGraw-Hill, New York, 1966.
- [47] Pytkowski, W., *The Dependence of Income of Small Farms Upon Their Area, Outlay and Capital Invested in Cows*, Biblioteka Pulawska, Warsaw, 1932.
- [48] Neyman, J. and Scott, E. L., 'Field Galaxies and Cluster Galaxies: Abundances of Morphological Types and Corresponding Luminosity Functions', in *Confrontation of Cosmological Theories with Observational Data* (ed. by M. S. Longair), D. Reidel Publishing Co., Dordrecht, 1974, pp. 129–140.
- [49] Neyman, J. and Puri, P., 'A Structural Model of Radiation Effects in Living Cells', *Proceedings U.S. Nat. Acad. Sci.* **73** (1976), 3360–3363.