ON THE FOUNDATIONS OF PROBABILITY AND STATISTICS, WITH SPECIAL REFERENCE TO RANDOMISATION THEORY*

by

J. N. Srivastava**
Harvard University and Cornell University

BU-531-M July, 1974

Acknowledgements and Abstract

This work arose out of certain comments on randomisation theory made by the author in various conferences during the past several years, particularly at the International Symposium on Statistical Design and Linear Models at Colorado State University, March 1973, and New York meetings of I.M.S. and other societies, December 1973. A great impetus to write was provided by Professor Federer. Long discussions with Professor Kiefer were very stimulating. I also had the privilege of brief discussions with many other persons, including among others, Professors Bose, Kempthorne, Barnard, Cochran and Ogawa. My thanks go to all of them.

We start with the randomisation theory, and consider the concepts on which it is based. Going deeper into the basis of this subject, we consider the problems of sampling and of the nature of probability, and consider some alternate conceptual approaches to statistical inference.

---

*This is a handout for a talk delivered at the Biometrics Unit, Cornell University on July 25, 1974, and is meant for only limited further circulation. The full paper will be prepared later. This work was partly supported by NSF Grant No. 30958X, at Colorado State University.

**On leave from Colorado State University.

I request the reader to adopt a "take it easy" attitude. The point is not that the present approaches are wrong. Rather, I am attempting to frankly examine the situation to see if some alternate conceptual approaches could be found which not only unite the present areas and constitute their foundation, but go beyond them (not only academically, but in some practically useful sense as well). As the reader will find, the attempt is as yet incomplete in every way.

1. Randomisation theory

Many workers, starting with Fisher, and later Neyman, Kempthorne, Wilk and others have considered the problems here. Recently Ogawa and Ikeda have provided many precise and mathematically rigorous results. Basically, the problem is this. Consider say, an ordinary randomised block design. Suppose the v treatments in each block are assigned randomly to the different units in that block. Let the permutation in the i'th block be $(\tau_{i1}, \cdots, \tau_{iv})$, $i = 1, \cdots, b$, and let the set of b blocks so obtained be denoted by D. Then the experiment is conducted using D. Let M be the mechanism (e.g. random numbers) using which the above permutations were obtained. Let $\{M\}$ denote the universe of possible designs D that can be generated by the mechanism M. If M corresponds to a completely random mechanism, as is ordinarily done, then each design $D \in \{M\}$ has the same probability of being adopted.

If in a given instance D_o is the design actually obtained using M, then D_o is used. However, once D_o is adopted, i.e. conditional to D_o, it is clear that the mechanism of generating the permutations such that under M_p, the probability that the design D is generated is p_D. Notice that p_D need not be necessarily equal. If $p_{D_o} > 0$, then it is possible that a person using M_p would obtain D_o,

and carry out the experiment using D_o . His experiment will in no way be different from that of the person who obtained D_o using M .

Now, in various randomisation theory results, both of Ogawa et al. and earlier workers, at a certain step the following "averaging" is performed:

If M was the mechanism used to obtain D_o, the D_o is embedded in $\{M\}$, and (a weighted) "averaging" is done over all possible D in $\{M\}$, the weights being the probabilities assigned to the various possible designs D by M .

A comment on the above approach is this: Suppose $D_o \in \{M_p\}$. Then using the above philosophy, we could also think of the universe $\{M_p\}$ in which D_o is embedded, and carry out the "averaging" using the probabilities p_D. Presumably, in general, "averaging" over $\{M\}$ and $\{M_p\}$ may give different results.

Actually, D_o is embedded in a large number of universes $\{M_1\}$, $\{M_2\}$, $\cdots$. The question is: Which universe among these should we choose for "averaging"?

2. Futuristic Inference

As a possible answer to the above question, I would like to make:

Suggestion 1: If it is decided that inference is to be made by embedding D_o in $\{M_i\}$ for some i, then embed D_o in $\{M_f\}$, where $\{M_f\}$ corresponds to a given universe of possible future "applications" of D_o . (It is understood that there may be many such universes in the future; it is suggested that for each case a separate embedding be considered. The inference may vary depending upon the nature of the universe).

To elaborate, I contend that if one decides to draw inferences from the results of an experiment, then one should have some idea of the situations where

he is going to use these inferences. This set of situations would supply the universe(s) required above.

If M_o is any particular randomisation mechanism considered in the last section, and many designs D_{01}, D_{02}, ··· are generated using M_o, and the corresponding experiments are performed, then embedding each in $\{M\}$ (and combining the results) would be 'valid', since the implied averaging over $\{M_o\}$ is physically taking place as a result of 'combining' the data from D_{01}, D_{02}, ··· . (Such a situation does occur sometimes; for example, in acceptance sampling.) On the other hand, it may be said that an experimenter should perform many experiments D_{01}, D_{02}, ··· (using the same M_o) since repetition is desirable and even necessary to 'confirm' the results. I submit, however, that experiments should not be repeated, that each experiment should be an <u>improvement</u> over the previous one, and that with improved and sequential experimentation, results should be confirmable faster than by a sequence of mere repetitions.

3. <u>Determinism, Conditioning, and Probability</u>

We noticed above that the central problem is that of "<u>conditioning</u>"; in other words, with respect to which universe $\{M^*\}$ should we look at a given D_o .

It seems to me that "conditioning" should play a very vital role in statistical problems; a number of mystical issues seem to be related to questions pertaining to "conditioning". When we make probabilistic statements, we must therefore emphasize the universe under consideration.

In terms of absolute concepts, I feel more like a determinist. In my opinion, it should be useful and enlightening to develop a deterministic foundation for statistics.

Under the deterministic attitude, the universe is all laid out (including the past and future). Thus, for an 'event' E, and 'trial' T, we have

(Absolute) Prob. (E will (did) occur at trial T) = 1, if T will be (was) held, and E will (did) occur.

= 0, if T ··· will not (did not) ···

= u, undefined, if T will not be (was not) held.

Thus the deterministic approach shows that we should (in order to have a useful theory) consider probability in the conditional way. Thus, probability should be looked upon more as a 'proportion' (i.e. a ratio of two frequencies, when all cases are 'equally likely') rather than a 'measure'; notice that it becomes a 'measure' only when the conditioning set is constant. The "ratio of frequencies" approach (i.e. one based on 'symmetry') is not the same as the "frequentist approach", where the conditioning set is a long (imaginery) series of trials. The connection between the two is well known. Subjective or personal probabilities also seem to be based on ratios of frequencies derived from past experience, using some (mental) averaging process.

Further comments on determinism and its impact will be made in the final draft of the paper.

4. Unbiasedness and Laws of Large Numbers

The laws of large numbers seem to come into grips with reality. However, we have to be careful with the concept of expected values. For example as is well known one can obtain unbiased estimators of each of the 128 parameters in a 2^7 factorial, by doing an experiment which says: Toss a coin. (1) If it falls heads, do not take any observation, (2) If it falls tails, select 2 treatments randomly out of the 2^7 possible ones, and observe the corresponding yields.

How "physically valid" are such estimators? This question brings in the notion of variance of the estimators. But the variance is obtained by conditioning the sample over an imaginary sequence of identical experiments. Should we consider such imaginary sequences as our 'conditioning sets'?

5. Inference and Design Conditioned on Actual Samples

Suppose the population of leaves of a tree is $N(\theta,1)$, and we have a sample $X_1, \cdots, X_n$ with mean $\bar{X}$. To test $H_o : \theta = \theta_o$, we use $y = n(\bar{X}-\theta_o)^2$ as a χ^2-variable. Then this population of χ^2-values is imaginary, except for one element. Saying that the value of y is significant at the $(1-\alpha)$ level refers to conditioning with respect to this imaginary population. Is such conditioning a conceptual necessity?

What about formulating the problem this way: How good is the above χ^2-test in the following sense (which is conditioned on the class of actual samples)? Suppose k statisticians (where, fortunately, k can be supposed to be large) each have a population $N(\theta_i,1)$, $i=1,\cdots,k$. Suppose each person has a sample of size n, and calculates χ^2 and physically carries out the test at $(1-\alpha)$ level. Approximately what proportion of these k statisticians would be rejecting <u>their</u> H_o when in fact it is true? Notice that imaginary populations are avoided in this formulation, at least in any explicit way.

Similarly, suppose a researcher has obtained a new design D^* which he considers 'better' than a given standard design D. The sense in which the word 'better' is generally used at present is that the 'variance' under D^* will be less than under D, the computations being made on two imaginary classes of possible observations that we would obtain under the two cases. A formulation of 'better' in the spirit of this section would be: If k experimenters (k large) use D^* rather than D, what

proportion of them would get estimates closer to the actual value of the parameter for their respective cases.

The questions under the present formulation can be easily answered if we bring in the imaginary populations. However, my point is: Can we conceptually bypass the imaginary populations completely?

Notice that in both the formulations, we are considering k users. Considering probabilities conditioned to these requires some knowledge of them. Thus in the first case, we could think of the k populations being describable by a frequency distribution of θ_o . Thus there is some shadow here of the mathematics of Bayesian-type of techniques.

6. Labels and Finite Populations

Coming back to randomization theory, we notice that if M^o generates D, and M^o corresponds to random assignments, then embedding D in $\{M^o\}$ corresponds to ignoring the names (or labels) of the units in the different blocks.

Godambe has raised the question of labels in sampling from finite populations. It seems to me that this is a basic issue in a big part of statistics.

The embedding question, and the futuristic or sample-oriented approaches suggested above apply to areas quite different from the randomisation theory of block-treatment designs. Thus as a coin is tossed and (TTTTH) obtained, the motivations could be (1) toss until an H is obtained, (2) toss 5 times, (3) toss until 2 tails and one head are obtained, and so on. Given the results of the experiment, there is no way to distinguish between the motivations. Clearly, the labels (here, the trial numbers) do have some importance. Similarly, in the Fisher's lady's tea tasting experiment, some interesting information may be lost by ignoring (a) trial

number, and (b) what kind of cup was offered at a particular trial, as Fisher does. For example, such information may be concerned with whether the lady had the ability to perceive correctly that a change in the kind of cup occurred whenever two consecutive cups of different kinds are given. (Of course, it seems to me that in his book, Fisher was concerned only with the smaller question of the lady's tea tasting ability.) With a given future application in view, Fisher's experiment could be embedded in a different universe.

The embedding question carries over to many nonparametric tests of the permutation type. Suppose, for example, there are 2n units $U = (1,2,\cdots,2n)$. Let $I = (i_1,\cdots,i_n)$ be a sutset of U . Let the value of a variable Y on the j'th unit be Y_j $(j=1,\cdots,2n)$. Suppose that to units in I, a treatment is applied which increases the response by an amount Δ . Then our observations will be $(x_1,\cdots,x_{2n})$ where $x_j = y_j$, $j \notin I$, $x_j = y_j + \Delta$, $j \in I$. Suppose we want to test $H_0 : \Delta = \Delta_0$. Let $z_j = x_j$, $j \notin I$, and $z_j = x_j - \Delta_0$, $j \in I$. Let $Z = (z_i,\cdots,z_{2n})$, $Z_K = \sum_{i\in K} z_i$, where K is any given subset of n units out of U, and let $d_K = |Z_K - Z_{\bar{K}}|$, where $\bar{K}$ is the subset of U complementary to K . Then a test for H_0 says: Look at the value of d_I within the collection of $\binom{2n}{n}$ numbers d_K; if H_0 is not true, then d_I will tend to be "large".

Usually, the subset I may be selected as a simple random sample. Then p_I, the probability of selecting I is $1/\binom{2n}{n}$. In general Z can be embedded in a universe where p_I have possibly different assigned values. Given I, Z is influenced only by I, not by the set of p_I for varying I . However, using the above philosophy, p_I should be incorporated at the inference stage.

I believe again that the futuristic inference approach provides an answer worth looking at.

7. Covariates, Optimal Design, and Randomisation

The remarks made so far occasionally pertained to randomisation theory, but not to randomisation. As more and more knowledge of the covariates becomes available, the design can be more and more optimised (w.r.t. reducing overall variance of treatment-contrasts), leading to more and more restricted randomisation. Thus, for example, in many situations, Fisher recommended Randomised Blocks (RB) over a Completely Randomised design, or a Latin Square over RB.

Heterogeneity may be of one or more of three types (1) Systematic variation or partially known type, like gradients in an agricultural field trial, (2) Systematic variation of a completely unknown type, and (3) Non-systematic variation, which would be akin to random variation. For (1), local control techniques, like blocking, in one or more directions, is advisable, leading therefore to lesser degree of randomisation. For (3), randomisation would serve no useful purpose. It is against (2), that randomisation (even of the restricted type) is considered to be helpful. However, to guard against (2), it appears to me, that certain "complex" systematic designs may have advantage over (restricted) randomisation. After all, often, random numbers are themselves obtained as a result of a "complex" but systematic (indeed, completely deterministic) process.

The subject of "complex systematic designs" will be discussed elsewhere, but until such designs are developed to a satisfactory extent, I would recommend (restricted) randomisation.