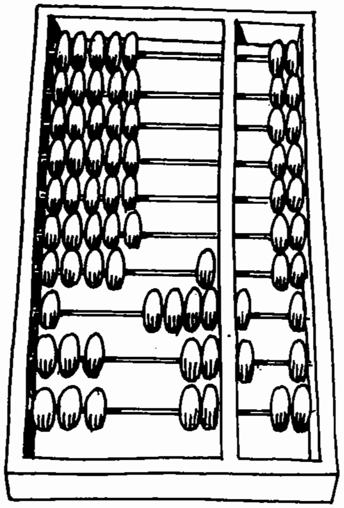

INFORMATION, SIMULATION, AND PRODUCTION: Some Applications of Statistics to Computing*

Mervin E. Muller *University of Wisconsin*†

Statistics can play a central role in fully utilizing the computer's power, economy, speed, and precision. The use of computers often depends upon the availability of statistical methodology to provide insight and direction in the specification of problems and in methods of solution. There are problems that, in spite of their social importance and in spite of the speed and power of computers, would be prohibitively expensive or time-consuming to solve without the use of statistical techniques to guide in the collection and analysis of data. We shall deal here with some of the ways statistics helps in the use of computers. Many other articles could be written to show how the availability of computers has made it possible to perform many kinds of statistical analyses and to use many kinds of statistical techniques that would have been impossible without a computer, because they would require excessive amounts of time, effort, or cost, or would be subject to excessive computational errors, but in general, we shall not treat these in this essay. Several essays in this volume base their analyses heavily on the use of the computer.

* The author is grateful to Alan G. Merten, Wayne Rayfield, and Peter J. Wolfe for their careful reading of the manuscript and helpful suggestions.

† Currently with the World Bank.

INFORMATION STORAGE AND RETRIEVAL

One important area of actual and potential use of statistics and computers is in the storage and retrieval of information. For example, with the increased mobility and size of our population, it is increasingly apparent how valuable it is to have fast access to one's medical history. One dimension of this problem is that patients and physicians are on the go, another is the explosion in the amount of medical literature and information now available for physicians; still another is the potential total size of medical files resulting from the size of our population. In some medical situations (arising from accidents, for example) it is exceedingly desirable to be able to locate rapidly a patient's medical history to determine possible dangers or side effects to the injured person if certain treatments are applied. Also, planned within the past few years is a pilot computer-based information storage and retrieval system for a National Poison Center. When this Center exists, any physician may call in, describe symptoms, and then be advised rapidly about the best possible treatments. The trouble is that as any of these medical files (or any other files of information) become large, the cost of storing and retrieving information can become large and the time required to find a desired item of information long. We need to arrange the information in storage so as to minimize these costs. Here is where statistical techniques can aid in providing criteria and methods of analysis for designing, implementing, and maintaining large computer-based information storage and retrieval systems which are both effective and economical.

Because we cannot usually afford to store all information items so that they can be retrieved in a short time, we are forced to take into account the demands for use of the items of information so that, in general, the most used items can be retrieved most quickly. But items vary over time in their frequency of use, so statisticians have developed theories for characterizing demands and estimating their fluctuations. This type of analysis can aid in determining where to store items of information because there will usually be varying demands for the use of information within a system. Hence we need ways to characterize such variations and then ways to use such analyses to design and evaluate the performance of an information system. Given such a design, other considerations, basically economic, control the type of computer hardware used and the actual speed of retrieval.

We shall now consider a very simplified problem in the organization of files to indicate some of the roles of statistics. Imagine that we must design an information retrieval system for information on symptoms and treatments of poisons. Suppose that the poison information file is to be kept in a sequential file. (To simplify the presentation, suppose the file contains information on only five types of poison.) We refer to the information on a particular poison as an information item, or more succinctly, as an item. Suppose that the five items (the poison information) are identified by labels A, B, C, D, and E. Usually there will be a different amount of information on each poison. For illustrative purposes, suppose also that the 5 items respectively contain 1, 2, 3, 4, and 5 information records so that the time it takes to scan item C is 3 time units because item C contains 3 records.

Once we decide on the order in which the items are to be placed in the file, we may assume that each time we have a request for some information on a particular poison we will start at the beginning of the information file and look at the label of each item until the desired one is found. (In place of a sequential file, we could consider other kinds of computer-information storage schemes so that we would not always have to start searching from the beginning of the file. These schemes present their own problems for determining the best file organization, but they will not be considered here. Such schemes offer attractive access speed but are generally more expensive to implement.) Suppose it has been possible to measure the relative frequency of demand for the five items as 1/10, 2/10, 4/10, 2/10, 1/10. Thus, items A and E are requested equally and least often, while item C is requested most often, 4 out of every 10 times a request is received.

What is a good way to arrange the file? We can call on the statistician to help us select a criterion of best arrangement. He might ask us to say whether or not in this case "best" implies that the average time to locate an item should be as small as possible or that the longest necessary search time should be as small as possible or maybe that the variability in search time should be as small as possible. Depending on the particular circumstances, each of these performance criteria could be the one of primary importance.

One criterion that is often selected is the minimum average access time, since it appears desirable to reduce the average time that someone must wait to obtain a desired answer to an inquiry. However, a file organized to minimize the average access time might cause some inquiries to the file to wait excessively long for a response. In some applications, minimizing the variability of waiting time for responses is considered to be most important. One measure of variability is the *range,* which measures the difference or spread between the longest and shortest wait. A different measure of variability is the *variance,* which measures the average squared deviation from the average access time.

In most cases, the file organization that is best with respect to one criterion will probably not be best with respect to any other criterion, as will be illustrated in an example. Therefore, the relative importance of the performance criteria must be taken into account when selecting one; this selection, in turn, will determine the file organization to be selected. For our example of five items, we might be prepared to use brute force, that is, to examine *all* the 120 possible arrangements of five items. If there were ten items in the file, a brute-force approach would require examination of over three million possible file arrangements, in fact, 3,628,800. With 10 items, brute force appears to be out of the question. However, such problems involving large numbers of possible arrangements are the sort that can be solved with the aid of statistical techniques.

The columns of Table 1 contain 10 illustrations of the 120 possible arrangements of the five items, and the last five rows contain some numerical values that indicate the statistical performance of the ten arrangements according to five criteria: average (mean) access time, variance of access time, minimum access time, maximum access time, range of access times. We see that orderings 1 and 2 possess average access time of 6.3 time units, the minimum for the ten chosen orderings and for all 120 possible arrangements as well. But the variance of the access time for arrangement 1 (15.21) is greater than for arrangement 2 (15.01). Arrangement 7 provides the ordering with the minimum variance (8.04) in the table and of all 120 possible arrangements, but a mean access time of 11.4 time units. Maybe we would prefer arrange-

TABLE 1. Ten of 120 Arrangements of Items A, B, C, D, E, Showing the Average Access Time, Variance, and Other Statistics

	ARRANGEMENT									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
First item	C	C	A	A	D	C	E	E	E	A
Second item	A	B	B	B	C	D	D	D	D	B
Third item	B	A	C	E	B	E	C	B	A	D
Fourth item	D	D	D	D	A	B	B	A	B	E
Fifth item	E	E	E	C	E	A	A	C	C	C
Statistics on access time										
Mean	6.3	6.3	6.6	9.9	7.9	8.1	11.4	11.7	11.7	9.3
Variance	15.21	15.01	15.24	29.09	9.09	23.89	8.04	10.41	10.61	29.61
Minimum	3	3	1	1	4	3	5	5	5	1
Maximum	15	15	15	15	15	15	15	15	15	15
Range	12	12	14	14	11	12	10	10	10	14

ment 5 because it has a reasonable (smaller than average) average access time and a relatively small variance.

For this example, the average of all 120 average access times is 9.0 time units. This average of all averages provides a basis for comparing how much better one can do in terms of average access by using the best organization instead of just leaving the file organization to chance and accepting the access performance associated with a randomly selected arrangement. Orderings 8 and 9 have the maximum average access time (both in the table and of the 120 orderings) of 11.7 time units and correspond to the reverse of orderings 1 and 2, respectively. Finally, ordering 10 has the maximum variance of access time of all 120 orderings, 29.61.

The complexity of analysis of data file organizations is much greater than the above example conveys. Statistical techniques used to design computer-based file organizations to have desirable storage and retrieval characteristics have many areas of applications, for example, keeping track of air-traffic patterns, voting records of elected public officers, dangerous drivers, and inventories of natural resources. One such technique, called the Monte Carlo method, will be described briefly later.

One of the oldest and most active uses of statistics in computer-based storage and retrieval of information involves dairy herds. The various dairy-herd improvement associations have record systems that contain extensive details about bulls, their offspring, and the associated milk production of the cows. Statistical techniques are used to help in determining what information should be kept in a file of this type so as to have an effective store of information.

MONTE CARLO APPLICATIONS

A class of statistical applications that was stimulated during World War II is identified as *Monte Carlo applications*. Some of the problems that needed solutions using these methods arose in the design of atomic reactors. Scientists had to estimate the behavior and interaction of elementary atomic particles so as to predict the expected termination site of the particles confined to the protective lead shieldings of the walls of the reactor. The term "Monte Carlo method" was selected because such methods use "computational statistical games" involving sampling and probability theory to provide understanding of the behavior of chance events involving atomic particles.

Insight into probability and statistics has been obtained throughout history by playing a particular game of chance many times. One use of the Monte Carlo approach would be to roll repeatedly a pair of dice, record the outcome of each roll, and then analyze all the outcomes in an attempt to determine if the dice were loaded. More generally, in the Monte Carlo method, we formulate and build statistical models and perform the necessary computations to generate samples from the models. When the results of these samples are analyzed, they provide useful estimates of the solution to a mathematical equation or equations.

There are several reasons why the Monte Carlo method is used. In some situations the real life process being studied has such complexities and uncertainties that even if a set of completely determined mathematical equations could be obtained to describe the process, no direct solution could be found, even if a computer with all its speed and accuracy were used. In other situations, no set of completely determined equations can be obtained at all.

The computer is made an effective tool in some of these situations by the application of statistical sampling techniques and mathematical models executed on computers. In this class of applications, some of the problems to be solved do not themselves directly involve statistics or probability. However, by use of statistical techniques involving sampling we can carry out on a computer a process whose statistical properties yield an estimate of the answer to the original nonstatistical problem. Thus the Monte Carlo method is an example of an interesting and important interaction between statistics and computers.

The Monte Carlo method requires developing statistical and computational procedures (called computing algorithms) that can be performed on computers to simulate various kinds of statistical or random behavior; for example, the computing algorithms provide the simulated probabilistic interaction of colliding atomic particles for a physics application, or a distribution of poker or bridge hands for one interested in understanding games of chance. It was partly through his desire to obtain the computational power needed for Monte Carlo–type analysis during World War II that one of the pioneers of the Monte Carlo method, John von Neumann, also became one of the most important pioneers of the modern digital computer. The term *simulation* is now often used in place of the term Monte Carlo method. Both terms are used to describe applications where statistical techniques directly influence the control and use of computers. In the file organization problem, we could obtain useful insight by performing Monte Carlo analysis or simulation to sample various orderings of the file and thus estimate the performance and characteristics of a near-optimal ordering.

COMPUTERS IN INDUSTRIAL PROCESSES

There are important industrial processes, in the chemical industry, for example, where the computer is recognized as a vital contributor. As an example, consider an industrial process from which some intermediate chemical B is desired. The process starts with product A, then yields B, and then yields C.

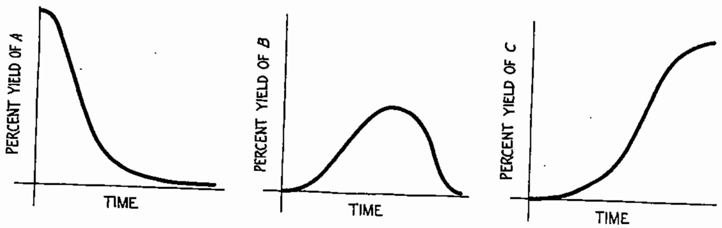

FIGURE 1
A chemical process that starts with chemical A has an intermediate yield of B and a final yield of C

These three parts of the process are shown separately in Figure 1 and then superimposed in Figure 2, so that the state of affairs at any time can be more easily seen.

The chemical engineer can be faced with considerable computational complexities in estimating the optimum time at which to begin extracting product B. Often he must use statistical techniques to determine a useful set of equations describing the chemical process. Of course, if B is extracted too early or too late in the process, only a small yield will be obtained. The engineer wants to find an interval around time *b* when the amount of B available is greatest. To characterize the process, he must collect data to verify the appropriateness of the equations describing the process and then collect data to estimate the parameters of the equations that are important in controlling the process. Without careful statistical analysis to determine *when* and *how* to collect data, not only will the computer analysis required to solve the complex and lengthy arithmetic be of questionable accuracy, but also the yield of the process may be substantially below what is required to have a useful and profitable product.

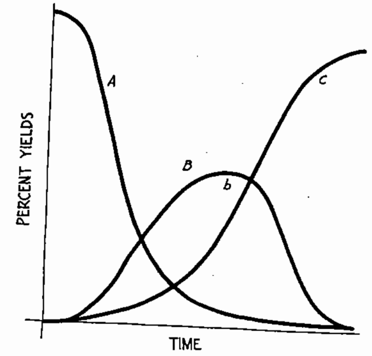

FIGURE 2
The process of Figure 1 on one time axis

REFERENCES

Alan G. Merten. 1970. "Some Quantitative Techniques for File Organization." University of Wisconsin Computing Center Technical Report No. 15.

Mervin E. Muller. 1969. "Statistics and Computers in Relation to Large Data Bases." R. C. Milton and J. A. Nelder, eds., *Statistical Computations,* New York: Academic. Pp. 87–176.