The Sampling of Records

By ROBERT E. PATTON, M.P.H.

Public health administrators have been making decisions on the basis of samples for many years. A sample taken from a water supply is examined and a decision is made about the condition of the entire water supply. The potency of a small sample of a vaccine is tested and from the results the potency of the entire lot is determined. A new immunization procedure is tested on a small group, and a decision is made as to whether the procedure should be put into general use or not. The sampling of records is a natural extension of these general sampling principles. Administrative decisions can be made from samples of records with at least as great a validity as those made from laboratory and clinic samples.

Restriction of the topic to the sampling of routine operating records—usually pieces of paper or cards—will simplify this discussion, excluding from consideration the special problems the statistician meets in taking a sample of persons by household interview.

We will assume that whenever we take a sample of a set of records we intend to measure or count something in the sample. From that measure or count we want to estimate the result we would have gotten if we had used all the records. That is the purpose of any sampling procedure—to estimate something about the whole by measuring or counting a part. It may be well to emphasize an obvious point. If the information necessary for a measurement is not available in a complete set of records that information cannot be put into the records by taking a sample. In other words, the investigator cannot find out from a sample more than he would from a full tabulation. He must realize, too, that all information in the records may not be usable. Frequently, material gathered as parallel or incidental information will not give valid answers.

The word "sample" needs defining. It is a mistake to call just any selection of items from a larger group a sample of that group. Deming in his book, "Some Theory of Sampling," uses the word "chunk" to represent any part of a total population (*1*). He restricts the word "sample" to mean a chunk selected in such a way that the reliability of estimates made from it can be determined. That is, results gotten from a sample can always be so stated that the reader can know the variability that is due to chance. He cannot always know the variability which must be ascribed to errors of response and interpretation, but he can measure the variability due to sampling.

This does not mean that a sample must always be selected randomly. There are other methods of sample selection for which the sampling error can be determined (*2*). The first 100 sheets of paper in a pile may be a sample of all the papers in that pile or it may not, depending on the circumstance. The question to ask is: Does each piece of paper have an equal chance of being selected by the method used? If this question can be answered affirmatively, we have a sample. In other words, we must know there is no more likelihood of one particular kind of paper being on top of the pile than of any other. If we can say and prove that any one

Mr. Patton, senior biostatistician of the division of local health services, New York State Department of Health, presented this paper June 18 at the Second Conference on Public Health Statistics, School of Public Health, University of Michigan, Ann Arbor.

of the sheets is as likely to be in the first 100 as any other, we have a sample when we take the first 100.

When should records be sampled? There is no one right answer that will apply to every circumstance. But there are two wrong answers—that records should never be sampled, or that they should always be sampled. It is not a "yes" or "no" proposition. Each specific set of circumstances must determine the answer.

Reasons for Sampling

There are three main reasons for sampling records. The first and most important in the field of public health, certainly, is to save money. When administrators are faced with the need for making the health dollar go as far as possible, the usefulness of record sampling as a money-saving procedure cannot be overlooked. Very often by tabulating only a small portion of a total set of records large sums of money can be saved.

A second reason for sampling records is to save time. The results of a sample tabulation can often be available much sooner than the results of a complete tabulation. In many public health situations timely information of slightly less accuracy is of more value than much more precise information months or years later. An excellent example of this use of sampling of records is the National Office of Vital Statistics monthly tabulation of deaths by cause for the country as a whole, based on a 10-percent sample. By this sampling device NOVS is able to get a national monthly tabulation of deaths by cause into print in the third month after the events occur. This represents a tremendous saving of the time that would be required if all the death certificates had to be coded and processed before the tabulations were made.

The third reason, and one that may seem almost paradoxical, is that more accurate results can often be obtained by using a sample of the records than by using all of them. For instance, if a survey of the quality of obstetrical care given to mothers of infants born in a State during a year were to be made from hospital records, it is possible that a more accurate evaluation could be made by doing a sample study than by looking at all the hospital records. In a sample study the number of records can be kept small enough so that the work can be done by one or two qualified obstetricians who can agree on standards of quality and on how to determine them from the hospital records. However, if the hospital records of all the mothers of children born in the State in a year had to be examined, it would probably be necessary to hire and train clerks for the work. Errors made by the relatively less well-trained clerks could far exceed the errors introduced by the sampling process.

A specific example may prove of interest. In New York State on July 1, 1951, a new birth certificate form was introduced in which the supplementary medical information section was changed radically. The existing supply of old forms in the field was not recalled, but requests for renewed supplies of certificates were filled with the new forms.

By January 1952, it was evident that a large percentage of the certificates arriving at the office of vital statistics in Albany were on new forms, and we were interested in knowing what the percentage for that month was. The certificates received in Albany are numbered and kept in numerical order by registration district and by date within each registration district. There were 12,910 births recorded in January 1952. Thus, the certificate numbers in January ran from 1 to 12,910. It was comparatively simple to pick one of the first 50 certificates at random and mark down whether it was on the old or the new form and then do the same for every fiftieth certificate after that. For instance, if the first one happened to be No. 27 we merely looked at certificate numbers 27, 77, 127, 177, and so forth. This gave us 258 certificates, or a 2-percent sample. In this sample we found 80.6 percent to be new certificates. This percentage has a standard deviation of 2.5 percent so that we are 95-percent sure that the true value lies between 75.6 and 85.6 percent. This was sufficiently accurate for our purposes. The whole job was done in about an hour. Examination of all the certificates would have taken a clerk at least a full day.

In order for any sample to be of value, it is necessary to have some idea of how closely the results of the tabulated sample agree with the results that would be gotten by tabulating all the data. In sampling, some variation is always introduced. If the sampling is done according to the principles of probability, the size of the sampling variation can be calculated. This is a back-handed definition of probability sampling. Turned around, it means that if the sample is drawn in such a way that the size of the sampling variation can be calculated, then it is a probability sample.

The Random Sample

A random sample is one type of probability sample. However, the word "random" is being used here in a very precise sense. It does not mean haphazard. It means that each element of a particular group or type has exactly the same chance of being selected in the sample as every other equivalent element and that there is no bias in the selection process. We can generalize this concept of randomness and say that the chance of an item falling into the sample need not be equal, but it must be a known chance. As long as this condition is present, we have a probability sample.

How can we pick a sample that is random? The common method is to drop each element into a hat and pick one element out at a time while blindfolded. This supposedly gives a random sample. It is essentially the method that was used to determine the order of induction of draftees prior to the beginning of World War II. Capsules which contained a slip of paper with a number on it were made up, put into a glass bowl, and selected one after another to determine the order in which men would be inducted. A study has shown that this procedure did not give a random result. There was too high a percentage of low numbers in the early draws and too high a percentage of high numbers in the later draws. This can best be explained by the fact that the capsules were carefully put in the bowl in order and evidently were not thoroughly mixed. That this classic example of random selection should have a bias in it is surprising. It does emphasize the fact that thorough mixing or shuffling is essential. Similar studies of bridge hands have shown that ordinary shuffling and cutting practices do not give a random distribution of hands. A better system of random selection is necessary.

Probably the best method of making a random selection from a series of elements is to number them and then to select the numbers of the elements to be used by a table of random numbers. On this table digits are arranged in a random order. Such a table is produced by using some mechanical device such as a numbered wheel, which is first tested to make sure that it does not have a bias in it. The results of successive spins are recorded in rows. This gives a series of random digits which can be used for many sampling problems. Random 2-digit numbers can be obtained by considering the digits in pairs. In a similar fashion random numbers with more digits can be obtained. The same starting point should not always be used. In fact, the starting point should be selected at random. Such tables are included in many statistics texts and collections of standard tables. When records are sampled, this is not a very practical method since it takes too long to number all the elements and make the selections. Something that will work much more easily and automatically than a table of random numbers is needed.

Systematic Sampling

Systematic sampling is such a method (*3*). By systematic sampling we mean the process of taking every fifth, tenth, or some other nth item on a list. The first item to take can be determined from a table of random numbers. Then one merely takes every nth item after that. It is the method described in the previous example about birth certificates. The question immediately arises: Is systematic sampling, random sampling? The answer probably is "not quite." However, it is probability sampling since, theoretically, the sampling variation can be determined. If we know something about the order in which the items are listed, we can, in general, tell whether the variation will be smaller or larger than the variation we would get by random sampling. In most practical situations, the variation is smaller in systematic than in random sampling.

This is particularly true if the list from which we are sampling is ordered by some characteristic which is correlated with the characteristic we are studying. When a systematic sample is taken from such an ordered list, we actually get a stratified sample. That is, we get a representative sample containing the correct number of elements from each portion of the list. If a list of last names were in alphabetical order we would get the same percentage of each letter of the alphabet in the sample as in the total list. Suppose the characteristic being measured is in some way associated with the last name, such as size of family. We would then have a smaller sampling variation for a systematic sample of the same size as a random sample.

However, systematic sampling may not work well if the list involves a cyclic pattern. Suppose we have a list of all the dwellings on a street. Suppose furthermore that all the houses on the street were built at one time by one builder and that he put exactly 12 houses on each block. If we took a systematic sample of 1/12 of those houses we could get either all corner houses or no corner houses in the sample. If we were studying some characteristic related to housing, such as income, the sample would be biased since people with higher incomes tend to live in houses situated on corners. Thus, a systematic sample which is representative for one characteristic may not be representative for another. A completely random sample on the other hand (which might have larger sampling errors) could be used for all characteristics.

But, if a cyclic factor such as the one which might have led to the sampling of corner houses is not present, there is no reason for using random instead of systematic sampling. And systematic sampling can save large amounts of time and money in the actual sampling process.

Sample of Medical Care Services

Two specific examples of how record sampling was actually used may help to explain some of these points. The bureau of public health economics at the University of Michigan was interested last year in studying the experience of Windsor Medical Services (*4*). This medical care plan, operating across the river from Detroit in Windsor, Ontario, provides physicians' service to about 100,000 individuals on a prepaid basis. A card record of all the service provided each individual who has a contract with the plan is maintained and filed alphabetically by the last name of the contractor. This contract may cover just the individual or it may cover him and some or all of his dependents. Thus, some cards have one name; most have several.

The study's primary aim was to describe and analyze the services received by individuals in the plan during a year's time. The facilities available did not permit a study of the records of all the 100,000 individuals enrolled in the plan. It was therefore decided to take a sample and to study the services received by the people in this sample. Preliminary calculations showed that the desired data could be obtained by using a sample of about 1,250 individuals. The results obtained from a random sample of 1,250 would have a desired accuracy 19 out of 20 times.

The problem, therefore, was the selection of 1,250 individuals from the 100,000 enrollees. Since some of the cards had more than one name on them, it was not satisfactory to select cards at random from the file. People who had individual contracts with the plan would have had a greater chance of being in the sample than people who were in the plan on a family contract. Since we wanted each individual to have an equal chance of being in the sample, some other scheme had to be used.

The cards on which the services were recorded constituted the only actual list of persons enrolled in the plan. It was not practical to number all these individuals, select 1,250 numbers from a table of random numbers, pick the cards for those individuals, and record their services for the year. The numbering job in itself was far beyond the facilities available.

We decided that systematic sampling was the most practical method to use. If we took every seventieth individual in this file of cards, starting at some random name among the first 70, we would then have a systematic sample of about 1,400 names, which would be satisfactory. However, even this was beyond our limited facilities. To get to every seventieth name in the file would have meant that every single name would have had to be counted, so a further compromise was decided on.

The cards with these names were kept in 30 file drawers. A random selection of $\frac{1}{5}$, or 6 of these file drawers, was made and every fourteenth name in those 6 drawers was selected. Since we had selected $\frac{1}{5}$ of the file drawers and were then taking $\frac{1}{14}$ of the individuals in those drawers, we were sampling at a rate of $\frac{1}{5} \times \frac{1}{14}$ or $\frac{1}{70}$.

By this procedure we got a sample of $\frac{1}{70}$ of all the names in the file, but we only had to count $\frac{1}{5}$ of the names in the file. We could have counted $\frac{1}{10}$ of the file and taken every seventh name, but we did not want more than one representative of the same family to be in the sample.

The results, which were obtained by this method and which could be checked against population values, were well within the expected random sampling variability. This is an indication, but only an indication, that the sampling procedure was adequate. There are two complications that should be mentioned. First, we were interested in all services received by subscribers during a specific year. The sampling was done about 3 months after the end of the year. The file of cards therefore contained individuals who had joined the plan during the year in which we were interested and in the 3 months following. Individuals who came into the plan during the year were accepted as part of the sample if they were selected. However, their date of entrance was noted on the record form, and in calculating the person-months of experience they were only credited with the actual time they were in the plan. If they had joined the plan during the 3 months after the period of study, their names were omitted if they fell into the sample. No other individual was substituted for them.

The second complication was that individuals left the plan during the study year. As they left, their cards were removed from the file and placed in an alphabetical discharge file. To insure that the sample would contain the proper proportion of these people, it was necessary that $\frac{1}{70}$ of all the persons who left the plan during the study year, or in the 3 months after the study year, should be in the sample so that we would have a record of the services they had received. Therefore, a procedure similar to that used on the main file was used on the discharge file. The record of services received by each individual selected was transferred to a special form which also contained spaces for such items as age, sex, and length of time in plan. From these forms cards were punched, and the desired results were easily obtained by machine tabulation.

Stratified Sample of Townships

Record sampling is also being used in a study now being conducted by the New York State Health Department. Fourteen of the 57 counties in upstate New York have full-time county health departments. The remaining 43 counties do not have organized county health departments. They are served by 15 State district health offices which coordinate the work of county nurses and other locally employed part-time and full-time health personnel. The director of the division of local health services wanted to know whether the people living in the rural and suburban areas were receiving more nursing, clinic, sanitation, or other types of service in the areas where there was a full-time county health department than where there was the combination of State and local services.

One practical way of measuring service was to use the routine operating records normally kept by the various health agencies. The job obviously called for some sort of sampling. There are slightly more than 2 million people in the rural and suburban areas being served by the district system and slightly over 1 million being served by full-time county health departments. To examine all the records pertaining to these 3 million people would obviously be impossible without a large staff and nearly unlimited resources. What was needed was a sample of all the records pertaining to a group of people who would be representative of the entire population being served by each administrative system.

Basically, there were two ways of cutting the records down to a manageable number. One would be to use the records of all the services which people received in a very short period of time—a day, for instance. This would involve many records in every office throughout the State. The variation in the services provided from day to day could be quite large. The second possibility was to take a sample of the people and study all the records applying to them for a longer period of time, say a year. This scheme was used.

The problem was to get a sample of the people in the rural and suburban areas served by each of the administrative systems. This was not simple; we do not have a list of all the people living in these areas. Even if we did, the problem of selecting the people from the list would not be trivial. We therefore decided to sample on an area basis. That is, we selected areas of the State and then looked for all the records which applied to services given to people living in those areas. The smallest unit of local government in upstate New York is the town. Each county is made up of a number of these townships, which are actually geographic areas and do not necessarily contain centers of population. The cities of the State are independent of the towns, but villages are included in townships. A compiled list of 927 towns included all the people that we wanted in the study. Of these, 717 were served by the State-local system of health administration and 210 by the full-time county system.

Next we selected a sample of towns whose residents would be a representative sample of all the residents of the rural and suburban areas of upstate New York served by each of the two administrative types of health systems. A completely random selection of the 717 and the 210 towns might or might not be the best sample to use since we could not be sure that we would not get too many large towns or, conversely, too many small towns. It seemed possible that residents in the small towns might receive less service. We wanted our two samples, one from each administrative system, to contain the same proportion of persons from large towns and from small towns as were actually in the areas served by the system.

What we were trying to find was the actual amount of service being provided during a given period by the two systems. We were not trying to find out whether one system gave more service than the other in comparable areas. The areas being served by the two systems are not comparable, and this sampling procedure was not intended to make them comparable. The samples were designed to be as representative as possible of each area as it actually exists.

Since we wanted to make sure that the final sample would contain the right proportion of persons from large and small towns, we decided to stratify the towns on the basis of size. This idea of stratification is really very simple. All it means is that the entire population being sampled is broken into smaller chunks and a sample is taken from each chunk. Each chunk should contain elements which are as nearly alike as possible. We listed each town by size, then divided them into groups which contained the same number of towns of the same size. The final sample consisted of one township from each of these groups.

One of the most important problems was the size of the sample. How many townships should be selected? Or, since one township was taken from each size group, how many size groups should there be? The factor that must determine the size of the sample, or how many townships to select, is always a compromise between the number one can afford to do and the number needed to be reasonably confident about the reliability of the results. To be 100-percent confident of the results of our study, we would have to take all the townships. The smaller percentage of confidence we are willing to accept, for a given range of our results, the smaller the sample we can use. The answer in this case was determined mainly by the available resources rather than by a desired reliability. This was partly due to the fact that very little initial information was available about what values could be expected for the variables we wanted to measure. For any type of probability sampling, the sampling variation that will occur for various size samples can be estimated. Estimates of the sampling variation can be computed on the basis of estimates of the values of the variables being measured and can be checked after the sampling is completed. We were interested in knowing what percentage of the people in these areas were receiving any service from the health agencies. When we started the study we did not know what the range of this variable would be. It could be as low as 1 percent or as high as 20 or 25 percent. It was a completely unknown quantity.

The same thing was true for nearly all the other important variables. It seemed best therefore to take as large a sample as the available personnel, funds, and time would permit. On this basis, we decided that about 20 townships might be handled. Since the average population of the townships is close to 3,000, we could expect a sample that would contain about 60,000 people.

Stratifying the townships on the basis of size insures the proper proportion of residents of large and small townships. It does not guarantee a good geographic distribution throughout the State. It seems quite likely that the variables we are interested in, such as amount of health service, will vary from county to county. Common sense would seem to tell us that we would have a better, more representative sample, if townships from many, rather than a few different sections of the State were presented in the sample. There is a technique which can be used to make the probabilities of such samples as large as possible. This has been called the use of controls beyond simple stratification in an article by Goodman and Kish. (*5*). This procedure was used in this study. It resulted in two very representative samples of the two areas of interest. Complete details of the sampling procedure will be published in the report of the study.

All routine and special reports should be discussed by the administrator and statistician jointly. One of the main points in this joint discussion should be: Do we have to use all the records or can we use a sample? This point must be discussed in terms of three factors: What is the purpose of the report? How accurate do the results have to be? What facilities are available to do the work? The facilities discussed must include equipment, personnel, funds, and time. Only after a balance has been struck between all these factors can an intelligent decision be made as to whether record sampling can be used for a particular report. The statistical and administrative questions are intimately related, and they must be answered jointly by the statistician and the administrator after open-minded discussion. There is no substitute for this teamwork approach to settling questions of record sampling.

REFERENCES

(*1*) Deming, William Edwards: Some theory of sampling. New York, John Wiley & Sons, Inc., 1950.

(*2*) Yates, Frank: Sampling methods for censuses and surveys. London, Charles Griffin & Co., 1949.

(*3*) Madow, Lillian H.: Systematic sampling and its relation to other sampling designs. J. Am. Stat. Assoc. 41: 204–217 (1946).

(*4*) Axelrod, S. J., and Patton, Robert E.: The use and abuse of prepaid comprehensive physicians' services. Am. J. Pub. Health 42: 566–574 (1952).

(*5*) Goodman, Roe, and Kish, Leslie: Controlled selection—A technique in probability sampling. J. Am. Stat. Assoc. 45: 350–372 (1950).

WHO Newsletter

Distribution of the *WHO Newsletter* with *Public Health Reports* is being discontinued. Individuals wishing to receive the *WHO Newsletter* may obtain it without charge from the World Health Organization Regional Office for the Americas, 1501 New Hampshire Avenue, N. W., Washington, D. C.