Targeted Surveillance

Written by Ziad Malaeb, Mathematical Statistician, PhD (A.B.D.),
Center for Animal Disease Information and Analysis, Risk Analysis Team

Disease does not occur at random. Rather, there are conditions under which disease is more likely to exist. Therefore, looking for disease or estimating its prevalence in a purely random manner would ignore these conditions and result in gross inefficiencies. In this monograph, we describe and propose a comprehensive sampling framework based on targeted sampling as a more efficient and effective way to find disease or estimate its prevalence. We make a distinction between statistical and biological confidence and demonstrate how to measure the overall surveillance confidence in a statistically rigorous and scientifically valid manner.

Surveillance activities to find disease or to assess its status based on random selections and testing of animals, especially when the prevalence of disease is very small in a large population, is extremely inefficient and often ineffective. Yet random sampling is the first sampling methodology that comes to mind when selecting a subset of animals from a large population for testing in a surveillance program. The main reasons for the popularity of random sampling are: (1) generalizability; (2) flexibility to choose *a priori* (before hand) a desired level of confidence and the ability to express it numerically, e.g., 95 percent; and (3) achieving statistical optimization in estimating parameters.

Generalizability

Random sampling—a specific form of probability sampling—is the number-one choice for selecting animals to be tested for disease for two primary reasons. First is the desire to infer from sample results to the entire population[1] at large. Second is the fact that only probability sampling, i.e., sampling where some random or probability device is used in the selection process, provides the logical basis for such an inference and allows the evaluation of its performance. However, if the purpose of sampling is solely to find disease in a large population, then random or probability sampling is not required. Instead, a more targeted sampling approach should be used. The problems with such an approach, however, are that one cannot make statistically valid statements about the presence or absence of disease in the general population at large, and no objectively measurable level of confidence can be attached.

Measurable Confidence

Another important reason for the popularity of random sampling in the selection and testing of animals is its flexibility in allowing one to decide *a priori* on a desired level of confidence, which can then be described with a number, e.g., 95 percent, that is convenient, valid, and meaningful. This convenience, however, has often caused a shift from reliance upon important epidemiological and subject-matter considerations to more statistical ones. As a result, statistical confidence has often been used as the sole measure of confidence in the sampling surveillance plan. Statistical confidence is only one part of

[1] A *population* here is taken to mean a large group of animals or herds susceptible to the disease in question and about which some information is desired. A *sample* is a part of the population under study, and the objective is to make inferences about the population from the sample.

the total confidence of the surveillance sampling activities. Relying more on statistical confidence in the context of finding disease can be misleading in the sense that it might give a false sense of security about the absence of disease. Again, this is particularly true when prevalence is very close to zero.

Statistical Optimalization of Estimated Parameters

The greatest majority of sampling activities entail estimating population parameters and testing hypotheses about them. Hence, most sampling techniques are designed to estimate parameters with estimators that have optimal statistical properties such as design-unbiasedness, consistency, and minimum variance. However, these statistical properties are irrelevant if the purpose of sampling is solely to find disease in a large population, not to estimate its prevalence or test hypotheses about it. Perhaps the only relevant aspects of probability sampling in the context of finding disease are sampling coverage and representativeness (i.e., freedom from selection bias), but only when one wishes to generalize sample results to the entire population at large.

Using random sampling to find disease when the prevalence (*p)* is near zero results in the problem becoming more probabilistic than statistical. This means there is little room for statistical manipulation and maneuvering. Also, none of the otherwise helpful sampling designs and statistical techniques can be used effectively to improve efficiency, incorporate biological confidence, or achieve scientific rigor. Furthermore, in pure probabilistic problems—those that obey certain rules of probability and can be completely addressed with probability laws only—confidence can be increased only by increasing the sample size. No matter the statistical maneuvers, when the prevalence *p* is small, one must increase the sample size to unrealistically high levels in order to achieve even a moderate level of statistical confidence. Often, important subject-matters are ignored and the problem is reduced to a mere calculation of sample size; this results in a measure of confidence that is purely statistical or, more correctly, probabilistic.

Questions of Surveillance

A few questions about the effectiveness of a surveillance effort are: (1) did you look in the right places; (2) did you collect enough samples; (3) were you able to recognize disease if you encountered it; (4) was disease distributed randomly and uniformly in the entire susceptible population, or were there subpopulations within it that disease might have been more or less likely to exist; and (5) how confident are you about your surveillance effort? Yet another issue in conducting surveillance over time is how to value the data collected, given continuous and dynamic changes in disease prevalence, population conditions, and other relevant circumstances (market conditions, for example).

What seems to be needed is a comprehensive sampling framework that measures both statistical and biological confidence objectively to provide progressively higher confidence as information is being gathered over time. We propose such a framework.

Sampling Strategy

Our proposed sampling strategy is based on targeted sampling. It uses weighted, stratified, adaptive sampling with probability proportional to risk of disease in a Bayesian

framework; it accounts for both statistical and biological confidence to give progressively higher confidence as information is gathered over time. This sampling strategy is efficient, effective, and relatively simple to implement. It is based primarily on epidemiologic and medical knowledge of the disease, rather than on pure statistical considerations. This strategy sequentially samples the entire population in an adaptive manner that takes into consideration other historical population information, including qualitative information, which produces progressively higher, measurable confidence. Stratification allows us to divide the population into smaller, non-overlapping subpopulations (strata) according to some desired criteria, such as the likelihood of disease existence. Adaptive sampling allows sequential, rather than one-time static sampling. And the Bayesian approach provides the framework for updating new information. We have also devised a mechanism that values data collected over time differently, giving more weight to the most current data.

This strategy effectively treats all the important issues of surveillance listed above, particularly those related to the ultimate question of measuring overall surveillance confidence. The National Surveillance Unit (NSU) has already used a form of this strategy in conducting its BSE surveillance. NSU is currently fine-tuning it to cover a wider range of objectives and apply to a host of diseases, rare and otherwise.

Please contact Ziad Malaeb at Ziad.A.Malaeb@aphis.usda.gov or (970) 494-7228 with any questions or comments regarding targeted surveillance.