

The Relation between Scientific Interpretation and Political Action: A Two-way Street

John C Bailar III
Professor, Department of Health Studies
And Harris School of Public Policy
University of Chicago
5841 S. Maryland Ave.
Chicago, IL 60637

Keynote address, Federal Committee on Statistical Methodology, November 15, 1999.

INTRODUCTION

The original meaning of the term "statistics" promises a great deal. The word is ultimately derived from the Latin word for "state", in the sense of a nation, and has been defined as the collection, analysis, interpretation and presentation of masses of numerical data. Ultimately, statistics in this sense is the study of political facts and figures.

This definition has little to do with how many professional statisticians use the word today. Although language changes, and fields of study change, sometimes rapidly, we have gone astray in making statistics a largely mathematical discipline, with emphasis on probability models and related matters. Our priorities are scrambled. Mathematical statistics is an important tool, but it is not the central tool of statistics. We must return our focus to the thoughtful analysis and interpretation of quantitative data.

THREE EXAMPLES

I will start with three examples. All are health-related, because I know health better than other fields, though the lessons I draw from these examples seem (from less concentrated study of other matters) to be quite general. I have chosen these examples because they have important policy implications and because they may be somewhat less familiar to this group of readers than other major statistical efforts such as a decennial census, monthly estimates of unemployment, or various aspects of our national accounts.

My first example is what has been called the "Gulf War Syndrome", or GWS. This condition has been said to affect many tens of thousands among the 700,000 U.S. veterans of the Gulf War of 1991-92. Gulf War veterans from other countries have expressed similar complaints. Symptoms attributed to the GWS include fatigue, malaise, headaches, bone and joint pain, difficulty sleeping, skin rashes, and many other things. However, each of these symptoms may occur rather commonly even in persons who did not serve in that war. Thus the question is immediately transformed, from whether veterans have these symptoms, to whether the frequency and/or severity of their symptoms is greater than in persons similarly situated but not deployed to the Gulf region in that conflict. In particular, there is much concern that this syndrome is a result of some kind of exposure --- perhaps chemical, including chemical warfare agents --- that was common in the Gulf area at the time these veterans were serving there. The most important policy implications have to do with diagnosis and correct treatment, prevention in possible future conflicts, and compensation for what may be a service-connected condition.

The Gulf War Syndrome has been examined in about 20 reports by different broadly-based scientific groups, all of which have concluded that there is no satisfactory evidence that the GWS is related to any unique physical exposure. One of these, issued recently by the RAND Corporation and reporting work done for the Department of Defense, cites some weak evidence that the syndrome might be related to a specific exposure in the Gulf theater, but also concludes that the evidence does not establish a physical cause for the problem. Given these expert opinions, why is there still so much uncertainty about the cause of the GWS and about our society's responses to it? Among the chief difficulties are: the lack of a unified definition of the syndrome (which seems to be whatever any research investigator --- or even any veteran --- wants to make it at the time), the lack of any suitable control group, and possible differences in reporting of

symptoms between persons who do versus those who do not believe they have the syndrome.

The problem of the Gulf War Syndrome is inherently statistical, because we need to know outcomes, and from those, determine risk factors and levels of risk for the occurrence of the GWS. Our understanding about these matters will have substantial impact on the way we deal with veterans of that conflict. I am quite skeptical that any specific chemical exposure is a cause in part because of evidence that U.S. veterans of every armed conflict that we have been in, at least as far back as the Civil War, have complained of similar sets of symptoms, though we do not have good estimates of their frequency. It further appears that veterans of other nations in other wars have had similar problems. A very active program of research is looking at many aspects of the GWS, and it is possible that evidence for some identifiable physical cause will emerge, but it is my present impression that the GWS is a version of a more general "war syndrome", or perhaps an even more general "stress syndrome".

My second example has to do with controlling the health effects of automotive emissions. There is solid evidence that very high levels of general air pollution sometimes seen in the past in many countries, particularly related to air that was stagnant over large cities over a long period of time, increased both morbidity and mortality, sometimes drastically. Many cities and countries have had much success in cleaning up the air, and today's questions center on whether today's reduced pollution levels still have some small, residual effects, and if so whether they merit further remedial action. Specifically, what we would like to know can be cast in a short series of questions:

1. What is getting into the air from various different sources?
2. How do those emissions translate into personal exposures (including personal variation such as for persons living near a major roadway or indoors versus outdoor locations)?
3. What effect do those exposures have on human morbidity and mortality?

Each of these is difficult to answer, and the linkages from one to the next are not easily established. The U.S. Environmental Protection Agency (EPA) has responsibility for air quality in the United States, and has constructed a series of rather elaborate statistical models to deal with various aspects of this, including automotive emissions (all classes of trucks as well as automobiles, and all types of fuels). Similar models have been developed in Germany and elsewhere. However, these models estimate emissions and then stop. We must take on faith that emissions translate into exposures, and that exposures translate into adverse health effects. But even emissions are difficult to measure, and there are numerous important matters for which one simply cannot obtain data. For example, few persons would want to have their own cars selected for full testing over a period of several days, and many would not want to have testing instruments on-board in their cars. We do not know how owners' perceptions of this car's possible problems may affect willingness to submit the car's for testing, so that test results may be correct for the cars tested but seriously biased for the population of cars on the road.

Or, we might want to know the effect on engine emissions of some change in the design of a carburetor, or a spark plug. The innovations can be put in test engines and their performance examined in detail on the laboratory test bed, or even in the field. But there is no way to tell, today, what might be coming out of an engine built today but still on the road twenty years from now, possibly after it has been poorly maintained or seriously abused.

We might want to estimate or predict the impact of a revised program of inspection and maintenance of vehicles. Many persons are now familiar with the routine, in which we may wait for hours for our cars to be briefly tested and certified as meeting regulatory standards for use in the road. In many states, testing includes sampling of the exhaust, and we blandly, or blindly, assume that those cars that fail to meet the test will either be repaired or taken out of service. There is good evidence that that is not true, at least in the U.S., and it seems that about two thirds of vehicles failing these state-mandated tests simply do not appear again in later testing records. Are the cars moved to another state with lesser or no standards? Are they converted to off road uses? Are some in fact repaired but not retested? How many are still in regular use, without the certification that the state requires? We do not know. EPA must model these and many other matters, and use the model to set standards that are known to be very costly, on the assumption that the unmeasured health benefits are likely to exceed the highly visible the costs.

Again, this matter is inherently and unavoidably statistical, since it deal with the analysis and interpretation of data, including data not available, that are related to a policy issue of considerable public importance.

A third example has to do with evaluating the impact of "block grants" from the U.S. federal government to states for the purpose of improving public health. Similar distributions are made in many other countries. Evaluation of such programs is difficult for several reasons. Some of these reasons are that funds from the central government may make up only a small part of what regional or local authorities are already putting into some program, that impacts for many kinds of public health measures (such as cancer prevention) are long-delayed in time, and that available data vary widely among the regions within a country, in scope, definitions, sample sizes, and quality. A committee of the U.S. National Academy of Sciences recently recommended four guidelines for the assessment of various measures that might be proposed to estimate the impacts of such programs:

1. Measures should be aimed at a specific objective and be result oriented.
2. Measures should be meaningful and understandable.
3. Data should be adequate to support the measure.
4. Measures should be valid, reliable, and responsive.

The set of measures that can meet all of these criteria is small. Thus, we must rely on limited, non-uniform, data of variable characteristics and uncertain quality. Again, this is a statistical issue, and it has much public importance with respect to these transfers of huge sums from central to regional and local governments.

LESSONS FROM THE EXAMPLES

These examples, and many others in health and in quite unrelated fields, have led me to recognize four features of statistical data available for use in almost any issue that has important policy implications. First is that the data are likely to be vast. One could fill many shelves, or even libraries, with material relevant to each of my three examples, and the scope of available material for some other problems would be much larger.

Second, the data tend to be highly complex, in the sense that the problem involves many different scientific and technical disciplines as well as non-scientific fields, and no one person can be expert in all of them. For example, full understanding of the Gulf War syndrome would require deep knowledge of military operations, generally, and in the Gulf Theater. The actual location of troops in the Gulf and their possible exposures; the health care systems for military personal and veterans (including medical health record systems) as well as medical care, more generally; toxicology; epidemiology; biostatistics; survey statistics (especially response biases); and many other things. Given that nobody can be expert in all of these simultaneously, how are we to deal with this complex situation? How can we identify all of the needed expertise, mobilize it, support it, and integrate it to solve an important problem?

Third, most of what is available will be of poor quality. Review of what is available on each of my three examples and many other topics has shown uniformly that a small amount of good work has been done, but that poor work is far more common, and that someone not expert in the field and appropriately critical could be seriously misled. Even organizing and getting through this mountain of literature, to determine what subset is worth detailed attention, is a daunting task.

Fourth, most of what we have is not going to be appropriate anyway. My best example of that is a fourth topic, related to both the GWS and automotive air pollution, which is the use of laboratory experiments to assess human hazards of exposure to chemicals. This is an important activity, and will remain important, because chemical substances are often considered for new uses, and we might hope that they will not be manufactured and made in quantity if they present serious and unrecognized hazards to human health. This means that our understanding of possible health risks must be derived from other sources, before extensive human exposure occurs. The best substitute in biological terms is animal testing. But such laboratory studies have grave limitations. Because of tight limits on the sizes of experiments, it is necessary to test small numbers of animals at high doses, and one must extrapolate from high to low exposures as well

as from animals to humans. There may be additional extrapolations from lifetime exposure to intermittent human exposure, or from one route of administration (such as food) to another (such as inhalation), as well as still other kinds of extrapolations. At the end, these uncertainties are compounded to a level where actual levels of risk to humans, or even whether there is any risk at all, maybe very uncertain. Repeated risk assessments of the same hazards quite commonly show differences of three orders of magnitude (one thousand-fold) or greater. I am not recommending that we abandon such risk assessments, which are generally the best we can do, but pointing to some inherent limitations on the precision of the estimates that can be produced. I firmly believe that, despite their problems, it's better to have a reliable summary of what is known, including uncertainties, than to just gaze into a crystal ball.

Consider these four features of data that I find are related to almost any policy issue – that they are vast, complex, of poor quality, and off target anyway – and see if you do not find echoes of them in your own fields of endeavor.

These considerations, based on my concern about the relation between statistics and policy, lead me to conclude that bias is almost invariably a matter of greater significance than randomness, at least in any context where the outcome matters very much. Statistics, in its original sense, is inherently an integrative discipline, but that sense has been largely abandoned. We may measure effects rather precisely in any one, small, well-designed research study, but there will inevitably be important differences among separate studies of the same phenomena, we should give substantial attention to what is known about related areas of the subject, and there is always scope for considerable interpretation.

REPRISE

If these four characteristics are common in the use of statistical concepts and data in public policy, what are we to do? The first step is recognition that there really is a problem. A few years ago, I saw a list of, as I recall, 31 major statistical agencies in the US Federal government. I was surprised and concerned to discover that 29 of the 31 Directors had little or no identifiable formal advanced training in statistics. If that remains true, we have a serious failure of the statistical profession in general to provide needed training and leadership in matters of great public importance. I do not suggest that any of the 29 directors were unqualified, and some indeed have become superb applied statisticians. However, they have had to acquire their skills from an unnecessarily narrow disciplinary perspective with, generally, a lot of on-the-job experience (that is, the error part of trial-and-error). Early, organized education about major statistical activities – profiting from the on-the-job experience of a broad range of earlier applied statisticians – would surely have been in their interest, and in the interest of all of us.

This should not surprise anyone. For many years, academic statistics has focused on mathematical concepts and computation. We do not attract (or attempt to attract) students who will take on major roles in the public sector as they move to their middle and later careers. We do not encourage students or junior faculty members to acquire supervised practical experience of the kind that one might get from internship in a statistical organization somewhere in government or the private sector. We do not generally recognize professional contributions outside the statistical literature. In short, persons whose primary professional identification is as statisticians have largely turned their backs on the large, real statistical problems that surround us. Instead, we devise elegant little probability models and apply them to elegant little sets of data that, too often, have no consequential significance.

If we continue to define statistics as the collection, analysis, interpretation, and presentation of masses of numerical data, we must recognize that the definition applies to many other professions as well. What is unique about statistics is its emphasis on general processes. The discipline of statistics should be focused on such things as asking the right questions in a way that can be answered, developing research protocols, maintaining and improving the quality of data, reducing massive detail to something that can be comprehended, integrating information of different kinds from different sources, applying probability models, and generalizing results, along with many other things.

As statisticians we may become experts in one or more of these, and such skills are portable. What one learns from one problem, in whatever field of endeavor, can readily be transferred to another. From this point of view, we many consider statistics to be what is left after we gather up the whole of science and technology and squeeze out all of the specifics of the disciplines of application, all of the specifics of individual problems, all of the specifics of research methods, and all of the specifics of individual study outcomes. What is left is the matrix of science itself, including the scientific method generally. That matrix can be applied to almost any issue where quantitative data may

be of some utility. If statistics, defined this way, is the science of any one thing, it is the science of the scientific method itself.

In short, our academic programs have failed us. We are not producing the kinds of graduates, at any level, that our nations need. If we truly believe in the glory and power of statistics, we must ask ourselves why we have so much less impact on world affairs than other professions with a similar technical or quantitative focus, such as economics, or geology, or high-energy physics, or applied mathematics. It is not just a matter of numbers of persons, or of bad luck. It is, in a deep sense, our own problem. We can solve it if we have the will.