National Research Council
Board on Mathematical Sciences

Symposium on Statistics in Science, Industry, and Public Policy

April 9, 1987
Washington, D.C.

National Academy Press
Washington, DC 1989

STATISTICAL CONCEPTS RELEVANT TO AIDS

Joel E. Cohen, Professor of Populations
Rockefeller University

INTRODUCTION OF DR. COHEN

Dr. Pyke: We turn now to the biological and medical sciences. Statistics has been involved there for years, possibly much longer than it has been involved in government statistics, and I am going to call directly on Dr. Joel Cohen, who is professor of populations at the Rockefeller University. He has been very active in the mathematical sciences and biological sciences and was a MacArthur Fellow. He will speak on statistical concepts relevant to AIDS.

Dr. Cohen: Thank you. I am pleased to be here. My title is "Statistical Concepts Relevant to AIDS," rather than the statistics of AIDS. If you want to find out about the statistics of AIDS, there is an article in the March 23, 1987 issue of a well-known scientific journal called New York Magazine which covers the numbers and describes fairly our understanding of AIDS.

My talk is going to be a talk in pure statistics. That means there won't be any data in it whatsoever. That reminds me of a story about a statistician and a physician who cared for AIDS patients. They were in jail the night before they were going to be executed for crimes that can only be imagined. The executioner came in and said to the statistician, "Do you have any last requests?" The statistician said, "I have been doing some work on statistical concepts relevant to AIDS, and I would like to give a seminar on my results, which are extremely interesting, to someone with a medical background who could evaluate them." He looked over at the doctor. The executioner said, "That is no problem. First thing tomorrow morning you will have your seminar." Then he turned to the doctor and said, "What is your last request?" The doctor's request was, "I want to be executed before the seminar."

Now, that story doesn't apply to me because I am not a statistician, and this talk is not for statistical professionals. I just study populations and use statistics to try to understand them. I try to develop tools, when the statisticians haven't provided them, that can then be used by statisticians.

There are many statistical concepts relevant to AIDS, and I am not going to talk about them all. I am going to talk briefly about a model of how an infectious disease spreads. This ridiculously simplified model has the great advantage of clarity. You can see what you need to measure. I am going to talk about diagnostic tests for the presence of the AIDS virus, since those are essential to knowing whether there is AIDS in a population. I am going to talk about survival time, and I am going to talk about surveys of sexual behavior because we know very little about people's normal sexual behavior.

What I am not going to talk about is uses of statistics in the evaluation of alternative treatments, such as in clinical trials. I am not going to talk about the use of statistics in evaluating or designing care by decision trees. I am not going to talk about the uses of statistics in laboratory research or experimental design and analysis.

BRANCHING PROCESSES AS A MODEL OF SPREAD OF INFECTION

Let us start by talking about the spread of infection in a population. There is a very simple model called a branching process. A newly-infected person infects a random number of what we may call offspring, that is, other newly-infected people. It is assumed in this model of a branching process that the probability rule that generates the number of offspring of any person is the same as the probability rule that generates the number of offspring of any other person, and that the offspring of one person are generated independently of the offspring of anybody else.

The spread of infection in the population depends largely on a single number, M, which is just the average number of offspring in repeated trials of this infection process.

> *"The spread of infection in the population depends largely on a single number M, which is just the average number of offspring in repeated trials of this infection process."*

If the mean offspring number M is greater than 1, it is likely that the infection will spread throughout the population. If M is less than 1, then it is certain that the infection will die out. M is an average over individuals of a product of several factors. The first factor is the probability for an individual that a sexual contact transmits a new infection. (I am ignoring the difference between transmission from males to females, females to males, males to males, or females to females.) A second factor is an individual's duration of infectivity. A person becomes infective at a certain time and remains infective for a period of time, which may end either when the person gets rid of the infection or when the person dies. A third factor is an individual's number of different sex partners per unit of time. So each individual has a period of exposure, a number of different partners per unit of time, and a probability of transmission at each contact. These are obviously central things one wants to know: transmission, duration, and number of sex partners. The average over individuals of the product of these three factors gives you M, and M predicts the spread or disappearance of infection in the population. You can get the same conclusions about spread of infection from much more complicated models.

> *"These are obviously central things one wants to know: transmission, duration, and number of sex partners."*

The branching process is a rudimentary, unrealistic model, but it focuses your attention on the main thing: on the average, how many new infections are created by one new infection? Before anyone swallows this branching process model as gospel, I should say what's wrong with it as a representation of reality. For starters, it ignores heterogeneity in the population. It assumes, for example, that the infected "offspring" distribution of a swinger in Manhattan is the same as that of a nun in Nebraska. Obviously there are major differences in behavior in different subpopulations, which greatly affect the likelihood of infection spreading into or out of those subpopulations. The model also ignores changes in time: changes in the frequency of "safe sex," for example, or in the infectivity, virulence, or genetic make-up of the virus. The mean number M of offspring is like a speedometer reading that tells you whether a car is traveling above or below the speed limit right now; because the world changes, today's value of M doesn't necessarily tell you tomorrow's or next year's.

In spite of its crudeness, the branching process is a useful model because it gives insight into what is happening at the moment. Unlike other models which may disguise equally crude assumptions in complex computer code or inaccessible mathematics, the branching process is transparent and raises questions that should be addressed by other models as well.

DIAGNOSTIC SCREENING

Now in order to talk about the probability of transmission, you have to talk about diagnosis. How do we measure whether there is an infection? A diagnostic test works in this way. You take a person who either does or does not have the infection, and you apply some kind of clinical procedure. The result is either positive or negative. In the case of AIDS, you take some blood, and you spin it in a centrifuge, and it separates into a solid part and a clear yellowish fluid part. That fluid is serum. You take the serum, and you put into it something which you hope will attach to antibodies which your body makes against the AIDS virus. The test looks for the chemical antibodies your body makes against the virus.

Two stages of recognition are involved here. The cells in your body that make antibodies have to look for the molecules on the virus, and your test has to look for the antibody molecules. In both cases there can be errors. Sometimes antibodies that look like they are against the virus are against other things in your body because there may be molecular similarities. A glove may fit more than one hand. So the antibody can fit more than one antigenic molecule. Secondly, your detector or test for that antibody is subject to error. The sensitivity of a diagnostic test is defined to be the probability that the test is positive given that the infection is present. It is the probability that if you have the infection, you are going to get a positive test. Obviously, the more

sensitive a test is the better, because you want to detect all the people who have the virus. The specificity is the probability, given that the person does not have the infection, that you get a negative test. It is the conditional probability of a negative test, given that the infection is absent.

The sensitivity and specificity of a diagnostic test do not depend on the prevalence of the infection in the population, that is, the fraction of the people who have the virus. You can just take a bunch of people, look at their results on these tests, and later observe whether they die of AIDS. You can measure sensitivity and specificity in a clinic or a laboratory without having to go out into the population.

When you are using a diagnostic test in a real population, you are less interested in the sensitivity and specificity than in the predictive value of the test. The predictive value of a positive test is the probability that the person has the virus given that he or she tests positive. The predictive value of a negative test is the probability that the person does not have the virus, given that he or she tests negative.

In most of the popular, and even many of the scientific, accounts of new diagnostic procedures for AIDS, there is no attention paid to sensitivity, specificity, or predictive values. How do you compute these predictive values? It turns out that the predictive values can be computed from the sensitivity and the specificity if you know the population's prevalence of disease. You use a theorem that is due to the Reverend Mr. Bayes, and we will come back to Mr. Bayes later.

> *"In most of the popular, and even many of the scientific, accounts of new diagnostic procedures for AIDS, there is no attention paid to sensitivity, specificity, or predictive values."*

Now, all of this is very abstract, but I would like to give you a numerical example of what this may imply which I hope will stimulate your intuition a little bit. As I said, there are no real data here. I am going to make up some numbers. Suppose you have a diagnostic test that is 99 percent sensitive. In every 100 people who have the infection, this test gives you a positive result for 99. Suppose also that it is 99 percent specific. For every 100 people who are free of the infection, it gives you a negative result for 99. Let us say the prevalence of infection in the population (like that of AIDS) is rare, say one percent. What is the predictive value of a positive test? Does anybody want to make a guess, leaving out the statisticians?

The answer is one-half. It means that if somebody tells you that some guy has a positive AIDS test, the probability is one-half that he really has the AIDS virus, and the probability is one-half that he doesn't.

If that is shocking to you, consider the pool of people who get positive test results; 99 percent of the infected people are going to be in that pool, but only one percent of the uninfected people are going to be in that pool. However, there are 99 times as many uninfected people as there are infected. So, the uninfected people contribute exactly as many people to the pool of people with positive results as do the infected people.

You may think you have a pretty good diagnostic procedure. When you apply it to a rare disease, you are going to get a lot of false positives. I don't think that all the people who use these diagnostic procedures are aware of this. Incidentally, the same problem (among others) afflicts the use of polygraphs to detect security risks. Until we really know what these sensitivity and specificity numbers are and until we have a good estimate of prevalence, we may be causing a lot of havoc. We should recognize the limits of our diagnostic tests.

Dr. Gerr (OFFICE OF NAVAL RESEARCH): What then is the predictive value of two successive positive tests?

Dr. Cohen: Suppose you take all the people who are positive after the first test. The prevalence of disease in that population is one-half. After another application of the diagnostic test, the predictive value of two positive results is 99 percent, provided you make one major assumption, namely, that the two tests are independent. If the two tests were strictly dependent, admittedly an extreme case, then the predictive value of two positive tests would be the same as the predictive value of one positive test, namely, one-half. Testing whether repeated applications of a given test to a given Joe Shmoe are independent requires a whole investigation in itself, which would ask a great deal of Joe Schmoe. If there exist empirical studies of the independence of different or repeated diagnostic tests for the AIDS virus, I am not aware of those studies. My conclusion: be careful with conclusions, or gather the facts needed to justify assumptions.

A last comment on diagnosis before we proceed to the next topic. We cannot evaluate a diagnostic test without attention to the costs of errors. Let me take two cases.

> *"We cannot evaluate a diagnostic test without attention to the costs of errors."*

Suppose I am screening blood in a blood bank. If I have a false positive, that means some blood tests positive even though there is no AIDS virus in it. I throw the blood away unnecessarily. But if I have a test that is a false negative, so there is AIDS virus in there, but I think it is okay, I am condemning someone who receives the blood to death. There you have one set of payoffs.

Now suppose I am applying for a job and somebody gives an AIDS test to me. If I don't have the AIDS virus and I test positive, my career may be severely damaged if I am refused the job. If I have AIDS and I get a negative test result I get the job, and I may be a source of risk for my fellow workers or I may not, depending on how I behave with my fellow workers. The point is that there are different sets of payoffs for different situations. I think we have to think much more conscientiously about sensitivity, specificity, and predictive values (positive and negative) as well as about payoffs before we decide whether and how to promote testing.

> *"I think we have to think much more conscientiously about sensitivity, specificity, and predictive values (positive and negative) as well as payoffs before we decide whether and how to promote testing."*

Dr. Koch (UNIVERSITY OF NORTH CAROLINA): Sometimes in cases like this one might want to adjust the background prevalence by noting that the subject being evaluated has certain risk factors in terms of behaviors and other exposures that would cause the prevalence for people like them to be higher than one in 100. That may then make the predictive value better.

Dr. Cohen: That is right. The U.S. Army, for example, screens all of its potential new recruits. They are getting about two percent positive according to the article on AIDS from New York Magazine. In other cases, among people applying for marriage licenses, tests give a much lower prevalence. Dr. Koch's point is that the predictive value depends on the prevalence in the subpopulation you are screening, and that is absolutely right. Nevertheless, an essential point is that there are only fallible tests. A battery of tests may have lower probabilities of error. But I think that discussions of policy related to AIDS screening need the numbers on predictive values as part of the discussion so they can consider the consequences of those numbers.

LIFE TABLES

So far, we have talked about branching processes and Bayes' Theorem. Now let me describe a third tool. Durations of infection, infectivity, and survival are summarized and compared using life tables. Here is the concept.

The life table describes the fraction of an original cohort surviving after some period of time. After a year you have say 90 percent surviving; after ten years you have 80 percent surviving, and so on. Life tables summarize survival and can be used to compare populations or treatments. You may treat one population with a drug and compare their survival with that of an untreated population.

Dr. Goldfield (NATIONAL RESEARCH COUNCIL): Is that a cohort of infected persons?

Dr. Cohen: You have asked the right question. What is this a cohort of? What is the beginning point? In this example, it is time since infection. The key problem is that you require a well-defined starting point. In discussions of survival of people with AIDS, I have read, for example, that middle-class patients are obviously getting better care because they survive two years, and the poor are getting worse care because they survive six months. But you don't know at what stage of the disease the diagnosis was originally made. You don't know whether there are differences in the starting point. So, as far as I am concerned, the comparison is meaningless without additional information. Let us standardize the starting point first. When do infection, infectivity, and disease begin? This is an issue that I think should be kept in mind in talking about survival.

> *"When do infection, infectivity, and disease begin? This is an issue that...should be kept in mind in talking about survival."*

RANDOMIZED RESPONSE IN SURVEYS

Finally, a last topic. How do you find out about people's sexual behavior while you protect their privacy? We have not had a study of American sexual behavior since Kinsey. But we have had a statistical invention since Kinsey called randomized response. It has never been used for a national survey of sexual behavior to my knowledge, and I would like to propose that it be used. Fiddler and Kleinknecht in 1977, and Dawes and Moore in 1978, suggested using randomized response to study sexual behavior and did so on a small scale. The need to extend the use to a national scale is now urgent.

> *"We have not had a study of American sexual behavior since Kinsey.... I would like to propose that [randomized response] be used."*

Here is how it works. Suppose I am interested in the fraction of people in this room who have had sex with a prostitute in the last month. Let me give you the simplest case of a randomized response. There are much more complicated ones, but I will start with the simplest. I call you up, and I say, "I want to find out what fraction of people had sex with a prostitute, but I don't want to invade your privacy. Please take a penny and flip it, but don't show it to me. If it comes up tails, and you did not have sex with a prostitute in the last month say 'No' to me. Under all other combinations, say 'Yes' to me. That is, say 'Yes' if it comes up heads and you did not have sex, or if it comes up heads and you did have sex, or if it comes up tails and you did have sex with a prostitute."

If nobody is having sex with prostitutes then about half the time I am going to get Yes's, and half the time I am going to get No's. Suppose that I get 20 percent No's. The only way I get a No is if the person did not have sex with a prostitute and got a tail on the coin toss. So, if I get 20 percent No's, then there must be another 20 percent who had no sex with prostitutes, because the coin does not know whether the person had sex with a prostitute. That means that 40 percent of the people were not having sex with prostitutes. That means, in turn, that 60 percent of the people were having sex with prostitutes. Equivalently, if I get 80 percent Yes's, I conclude that 60 percent of the people were having sex with prostitutes.

If you say "Yes" to me, I will not have the slightest idea whether you had sex with a prostitute and I don't really care, because I want to know the frequency for the population. But if you say "No" (honestly), I know that you didn't have sex with a prostitute. "Yes" and "No" are not symmetric in this version of randomized response.

Participant: There is a probability that a coin will not give tails exactly 50 percent of the time.

Dr. Cohen: Good point. In a sample of 100 coin tosses, you are not going to get exactly 50 tails. You have to allow for that variability when you estimate the uncertainty of your final results.

Participant: What if you want to measure the frequency of a rare phenomenon?

Dr. Cohen: Then you use a biased coin that is tuned for the estimated probability in advance. You can make a smarter coin or other randomization device that will save you trouble. Randomized responses of this kind have actually been used in practice to estimate frequencies of tax evasion, drug use, illegal telephones, and illegal abortions (when they were illegal). There are versions for mail, for telephone, for personal interviews, and for quantitative questions like, "How many sex partners did you have in the last month?"

FOUR CENTURIES OF STATISTICS

In conclusion, let me review where the concepts I've been describing came from. Branching processes were invented by Bienayme in 1845 and by Galton and Watson in 1873. Galton was related to Darwin. They were reinvented by some physicists and a statistician in the 20th century, but really they are 19th century objects. The rule for getting predictive values of diagnostic tests based on Bayes' Theorem was published in 1763, two years after Bayes died. The life table was invented by two Englishmen, Graunt in 1662 studying the London bills of mortality, and possibly by Petty in 1683. The idea of randomized response was invented by Stanley Warner in 1965. What I would like you to see is that we have a problem with AIDS today. We can draw on four centuries, the 20th century, the 19th century, the 18th century, and the 17th century, four centuries of the development of statistical ideas, to deal with that problem.

"What I would like you to see is that we have a problem with AIDS today. We can draw on four centuries....of the development of statistical ideas to deal with that problem."

According to Salomon Bochner, "Statistics founded on probability is perhaps the most exclusive characteristic of our civilization since 1600; and it would be difficult to find even a trace of it anywhere before." Even if Bochner was a Princeton professor, this is probably true.

DISCUSSION

Dr. Polking (NATIONAL SCIENCE FOUNDATION): I have a question about the randomized response method. How do you convince your survey population that the whole situation really does protect their confidentiality? It seems to me that this is a real problem with that kind of survey.

Dr. Cohen: I agree completely. It is a matter of empirical research with psychologists to find out what is the most persuasive way to present questions with randomized response so that people accept that their confidentiality is being protected. There have been field trials of randomized response. We have to continue to learn, just as we have had to learn about the phrasing, ordering, and presentation of ordinary direct questions. It is a matter for research.

Dr. Horvitz (RESEARCH TRIANGLE INSTITUTE): I had some earlier experience with randomized response. We did a study around 1968 on illegal abortion. At that time the Supreme Court decision had not been made. The study was among women in urban places in North Carolina, and we had some reasonably good success with it, at least after the fact. From the study we estimated there was about 23 percent pregnancy wastage from illegal abortion. I think current statistics put the figure somewhere around 27 percent from abortion.

We also tried to validate the study at a time when we felt that having an illegitimate birth was a stigma. We took a sample of births, both illegitimate and legitimate, and we inquired about whether there had been a birth in a household where we had identified from the birth certificate at that address that the woman was not married. In comparison with the randomized response approach, we were able to get rather good estimates of the actual proportion. We didn't design our study in a way that would have determined whether or not we could have gotten the answer anyway by direct question. That was a bit unfortunate. We also did a study with people who were known to have been arrested for driving under the influence, and here we got no success whatsoever.

From my experience, randomized response is very worthwhile trying, but it does need a lot of research on the psychology of the public, and in what instances they would respond or not respond. It is worth developing.