A METHOD FOR ANALYZING LONGITUDINAL OBSERVATIONS ON INDIVIDUALS IN THE FRAMINGHAM HEART STUDY

Harold A. Kahn, National Heart Institute, National Institutes of Health

The Framingham Study was established in Framingham, Massachusetts in 1948 to measure the development of cardiovascular disease in a sample of the town population aged 30-59 and to determine factors associated with disease onset. Data are obtained primarily from a scheduled biennial examination conducted in a clinic maintained especially for this study. Although the examinations are intended to be at 2 year intervals, they are often longer or shorter by several months and in some cases are very much longer whenever one or more complete examination cycles is missed. Additional description of the Framingham Study together with some of the substantive findings to date can be found in an article by Kannel, W.B., et al., in ANNALS OF INTERNAL MEDICINE 55:33 July 1961. For the present let's concentrate on the repeat examination feature of the study.

Consider where we are today. Approximately 4400 individuals have had 5 or more examinations, about 3400 of these have had 6 or more and about 900 have had 7. However, analyses up-to-date have been based almost entirely upon associating the characteristics found on the first examination with subsequent outcome regarding disease. What we are trying to do now is to associate with subsequent outcome, all of the measurements made on a variable over the entire study period. As a specific example suppose that we consider the measurement of total serum cholesterol on successive examinations. Two hypothetical individuals might have readings as follows:

"A" 220-240-260-280-300 (died of CHD)

"B" 280-260-240-220-200-180-170 (still living)

Our interest is in finding the association between serum cholesterol level and development of coronary heart disease (CHD). Obviously we get somewhat different views of this association if we look at all of the measurements made or if we look at only the first examination results. It seems reasonable to presume that using all rather than part of the data available will lead to a better description of the relationship. The question is how to do this. For convenience we will refer to cholesterol measurements throughout the rest of the paper although obviously the procedures discussed are applicable to other variables and with appropriate modification to attributes as well.

One approach would be to consider in a single unit all the cholesterol measurements made on an individual. He could then be classified according to various functions of these data. However, since all the measurements are used, the observation time after establishment of the individuals classification category is very small. In effect, we would have "used up" the time under study to establish a classification and would have little time left over in which to see what is the subsequent incidence of disease associated with various classifications. It would, of course, be possible to wait for enough time to elapse following the series of cholesterol measurements so as to observe incidence rates associated with various categories. Such a waiting period implies not using later measurements to up-date the individuals classification, since to do so would bring us right back to practically all classification and almost no subsequent observation time.

There are two possible sub-divisions of the method suggested above. (1) restrict the analysis to just those who have not experienced the event being studied over the entire period during which cholesterol measurements were made or (2) also include any individuals who are no longer under observation at the end of the measurement period because the event being studied did occur at sometime during the measurement period. In case (1),we would be limiting our study to that subset which "survived" the measurement period and could not properly apply our findings to the total population. Case (2) has a different defect which can be illustrated by considering two individuals with cholesterol measurements as follows: "C" - 200-200-200 (Died of CHD); "D" - 200-200-200-220-240-260-280. If we classify the trend of measurements for each individual as a unit we shall categorize individual "C" as one showing a level trend and "D" as one showing an upward trend. We would then associate the unfavorable outcome observed for "C" with level trend experience and await subsequent developments in order to associate an outcome with the upward trend observed for "D". The method of treating the individual as a unit is not satisfactory in that individual "D" did have level trend experience for a period which was identical to that for "C". However no event occurred to "D" immediately following this level trend experience and because his total experience is considered as a unit the method makes no provision for recording what is in fact the truth: that two individuals showed level trend experience for a similar length of time and that only one of these had an event. A method which counts cholesterol classification categories which are followed by an event, but doesn't count an identically classified category which is not followed by an event, cannot be recommended. The method lacks a clear way of determining how many years of observation are associated with a particular classification. We could make a

reasonable approach toward answering this question if we considered the total classification period in segments and assigned various segments to those cholesterol categories with which they would be identified if an event occurred during the segment, but this would lead us away from treating the classification measures on an individual as a unit. We turn now to description of an alternative approach which is not restricted in this way and which uses all of the classification data and all of the observations regarding occurrence of events.

This approach depends upon breaking up the entire observation for each individual into periods beginning with each examination date and extending up to, but not including the next examination date. Each of these observation periods is classified according to cholesterol category as of the start of the period. However, there is no reason to exclude data on cholesterol classification available from prior examinations. We do, in fact, use various functions of all of the cholesterol measurements made up to and including the examination which defines the beginning of the observation period. Thus, we can categorize an observation period according to cholesterol measurements such as the value at the beginning of the period, the average, the maximum, the standard deviation or the slope using all examination data up to and including the beginning of the period.

We have defined an observation period as the interval from one examination date to the next one (with appropriate modification when there is no "next" examination because of death, or loss from observation). Since the longer the period the greater the chance that a new case of disease may occur in it, we relate the occurrence or non-occurrence of the event being measured to the length of the observation period within which it had some chance to happen.

At this point it may help to introduce the following notation:

Let L_{ij} = length of observation period following jth exam on ith individual

X_{ij} = serum cholesterol measurement as of the jth exam for ith individual

Y_{ij} = 1 or 0 according to whether the event being measured (say the occurrence of myocardial infarction) occurred or not to the ith individual during the observation period following jth exam.

We will compute the statistic $\frac{\Sigma\Sigma Y_{ij}}{\Sigma\Sigma L_{ij}}$ for which the corresponding X_{ij} are all in a specified cholesterol classification category. For example the cholesterol category might be one in which the average value of all measurements made prior to and including the beginning of the interval is greater than 260 mg%. The numerator would then be the sum of the cases occurring in intervals classified as above. The denominator would be the total length in years of the observation periods so classified. The statistic $\frac{\Sigma\Sigma Y_{ij}}{\Sigma\Sigma L_{ij}}$ is an incidence rate per person year which we can associate with the particular cholesterol classification for which it was computed.

The entire procedure is essentially one of pooling together data from different individuals, as well as from different time periods for the same individual for which the common element is cholesterol status according to some specified definition. For instance an incidence rate associated with a cholesterol average of less than 200 mg% might be made up of data from 2 consecutive time periods for individual A plus two non-consecutive intervals from individual B, etc. Obviously such a wholesale hashing together of available data would not be very helpful without appropriate restraints. One of these is to combine data in age-sex specific groupings so that there is a reasonable approximation to homogeneity of combined experience. Ten year groupings are probably adequate in this regard. Another restriction is to omit observation periods for which the risk of developing an event is zero. Thus all observation periods subsequent to the occurrence of an event are excluded from the calculation of an incidence rate of that event.

We have defined the length of the observation interval according to the following rules: If there is a subsequent examination date, the interval ends on that date. If there is no subsequent examination we have considered the interval indeterminate and have not used it for events such as angina pectoris or hypertension which are not likely to be discovered except by examination. For events such as myocardial infarction or death which have a high probability of detection in the study mechanism, the observation interval for individuals still living in Framingham, ends on the anniversary date following an event (if there is one) or the cut-off date of the tabulations whichever is earlier. For those who have moved out of town the observation period ends on the estimated moving date.

In order to compute age-specific incidence rates it is necessary to assign observation periods and events to age classes such as 45-54, 55-64, etc. If the entire observation period is within a single age class, there is no problem. If it should overlap two, the length of observation period is assigned on a pro-rata basis to each except that the older age class assignment is cancelled if the event being measured took place in the younger age class. This is in keeping with the rule to avoid counting observation periods for which the risk of an event is zero.

Events such as myocardial infarction or death can be easily allocated to correct age classes. Events such as angina pectoris or hypertension are arbitrarily assigned to the mid-point of the interval in which they were discovered.

The incidence rates for different cholesterol categories that result from applying the above definitions have an unusual property in that the same individual may be contributing his experience to the incidence rates for both high and low cholesterol. Of course such incidence rates are not independent and significance tests of the difference in rates between groups are complicated thereby. Fortunately, the covariance between rates of the size we are concerned with here (say .01 per person year) is sufficiently small in comparison to the variance that we may safely neglect it. Derivation of the variances and covariances of the incidence rates for these various categories is presented in the following appendix.

The most important feature of the incidence rates described is that they are able to use all the cholesterol information available in relating cholesterol category to disease outcome. Given a long term prospective study with sequential measurements such as the Framingham Heart Study it should be possible in time to compute incidence rates associated with cholesterol categories such as the following:1/

the most recent value
the mean
the variance
the maximum
the slope, or
the length of time during which all measurements have exceeded a specified value

APPENDIX

All of the following should be understood as referring to rates for a particular classification category for cholesterol and to a particular age-sex group.

As before: Y_{ij} = 1 or 0 depending on whether the event being measured happened or not to the ith individual during the observation period following the jth exam

L_{ij} = length of the observation period following the jth exam for ith individual (in years)

1/ Computer programming of the Framingham data began in July 1961.

$p = \dfrac{\Sigma\Sigma Y_{ij}}{\Sigma\Sigma L_{ij}}$ is computed for ij in which the classification variable (cholesterol) is in a specified category and for which the L_{ij} are in a particular age-sex group.

Among all the possible samples which could have been chosen from the Framingham Adult population we are going to restrict our interest to that subset with the same amount of observation time in each classification category as the actual sample we have. This seems like a reasonable restriction which both simplifies and improves our results since variance associated with variation in observation period is somewhat extraneous to our interest.

We shall also postulate that incidence rates are constant for all observation periods within any specific age, sex and cholesterol bracket. Actually this is not true, but it is probably sufficiently close for our purposes.

Designating P as the true incidence rate, we have,

$$E(Y_{ij}) = PL_{ij}$$

where the P and L_{ij} are of such size that their product lies between 0 and 1.

$$V[p] = \frac{1}{\left[\Sigma\Sigma L_{ij}\right]^2} \Sigma\Sigma E\left[Y_{ij} - PL_{ij}\right]^2$$

$$= \frac{1}{\left[\Sigma\Sigma L_{ij}\right]^2} \Sigma\Sigma\left[E\, Y_{ij}^2 - P^2 L_{ij}^2\right]$$

$$= \frac{1}{\left[\Sigma\Sigma L_{ij}\right]^2} \Sigma\Sigma\left[P\, L_{ij} - P^2 L_{ij}^2\right]$$

$$= \frac{P}{\left[\Sigma\Sigma L_{ij}\right]^2} \left[\Sigma\Sigma L_{ij} - P\, \Sigma\Sigma L_{ij}^2\right]$$

Of course, we do not know P and would substitute the sample observation p in place of it. A necessary qualification on this formula is that it is appropriate only if $\Sigma\Sigma L_{ij} > P\, \Sigma\Sigma L_{ij}^2$. In the Framingham study with most observation periods 2 years long and annual rates of about .01, $\Sigma\Sigma L_{ij}$

will generally be about 25 times as big as $P\,\Sigma\Sigma L_{ij}^2$. The previously stated restriction that $0 < PL_{ij} < 1$ is sufficient to assure that the variance of p will always be positive.

Although the preceding form is the one used in computing, it is helpful to get an idea of what this variance amounts to under the assumption that all L_{ij} are of equal length say L. This means that $\Sigma\Sigma L_{ij} = mL$ where m is the number of observation periods contributing data to the incidence rate under consideration. Then:

$$V\left[P\right] = \frac{P}{m^2 L^2}\left[mL - P\,mL^2\right]$$

$$= \frac{P}{mL}\left[1 - P\,L\right]$$

which is noticeably different from the variance of a binomial variable in that the denominator is person years rather than persons and the numerator is not of the form P(1-P) but P(1-PL).

The coefficient of variation is insensitive to the units in which the observation is measured, i.e., changing from person years to person months will divide the rate by 12 and the variance by 144.

To find the covariance between p_1 and p_2 where p_1 is the sample incidence rate for an age-sex cholesterol category and p_2 is a similar rate but for a different cholesterol category, we recall that:

$p_1 = \frac{\Sigma\Sigma Y_{ij}}{\Sigma\Sigma L_{ij}}$ for cholesterol category 1 and

$p_2 = \frac{\Sigma\Sigma Y_{ij}}{\Sigma\Sigma L_{ij}}$ for cholesterol category 2. To help distinguish cholesterol categories we shall introduce "1" and "2" into the notation preceding the subscript for individuals. The two rates will in general be made up from data for different sets of individuals and it will also help to change notation for individuals in group "2" from i to k. Since the covariance between rates depends upon the same individual contributing to both rates and isn't affected by the number of observation segments into which that individuals experience is divided, we shall simplify our notation by dropping reference to subscript j for observation periods. However it should be understood that all references to an individual are for the sum of his observation periods appropriate to the cholesterol and age classification being considered. Thus we have:

$$p_1 = \frac{\sum_{i=1}^{n_1} Y_{1i}}{\sum_{i=1}^{n_1} L_{1i}} \quad \text{and} \quad p_2 = \frac{\sum_{k=1}^{n_2} Y_{2k}}{\sum_{k=1}^{n_2} L_{2k}}$$

The covariance of p'_1 and p'_2 is:

$$CV\left(p_1 p_2\right) = E\left[p_1 - E(p_1)\right]\left[p_2 - E(p_2)\right]$$

$$= E\,p_1 p_2 - E\,p_1\,E\,p_2$$

$$\text{where } p_1 p_2 = \frac{\left(\sum_{i=1}^{n_1} Y_{1i}\right)\left(\sum_{k=1}^{n_2} Y_{2k}\right)}{\left(\sum_{i=1}^{n_1} L_{1i}\right)\left(\sum_{k=1}^{n_2} L_{2k}\right)}$$

In order to see how large the covariance might be under extreme conditions we will assume that $n_1 = n_2 = n$ which would only be the case if all individuals contributed at least some of their experience to each of the rates. We can then break up the numerator into two terms.

$$p_1 p_2 = \frac{\sum_{i=k}^{n} \left(Y_{1i} Y_{2k}\right) + \sum\sum_{i \neq k}^{n} \left(Y_{1i} Y_{2k}\right)}{\left(\sum_{i=1}^{n} L_{1i}\right)\left(\sum_{k=1}^{n} L_{2k}\right)}$$

The first term in the numerator is the product of $Y_1 Y_2$ values for the same individuals. Since the only possibilities here are 0.0, 0.1 and 1.0 we can substitute zero for this term. 1.1 is not possible, nor is it even defined, because we do not include observation periods after an event.

$$E\,p_1 p_2 = \frac{\sum\sum_{i \neq k} E\left(Y_{1i} Y_{2k}\right)}{\left(\Sigma L_{1i}\right)\left(\Sigma L_{2k}\right)}$$

Since $i \neq k$ the Y_1 and Y_2 values are for different persons and therefore independent. Thus we get:

$$= \frac{\sum\sum_{i \neq k} P_1 L_{1i} P_2 L_{2k}}{\left(\Sigma L_{1i}\right)\left(\Sigma L_{2k}\right)}$$

$$= \frac{P_1 P_2}{\left(\Sigma L_{1i}\right)\left(\Sigma L_{2k}\right)} \frac{\left[\left(\Sigma L_{1i}\right)\left(\Sigma L_{2k}\right) - \Sigma_{i=k} L_{1i} L_{2k}\right]}{}$$

$$= P_1P_2 - P_1P_2 \frac{\sum_{i=k} L_{1i} L_{2k}}{(\Sigma L_{1i})(\Sigma L_{2k})}$$

$$CV(p_1p_2) = P_1P_2 - P_1P_2 \frac{\sum_{i=k} L_{1i}L_{2k}}{(\Sigma L_{1i})(\Sigma L_{2k})} - P_1P_2$$

$$= - \frac{P_1P_2 \sum_{i=k}^{n} L_{1i}L_{2k}}{(\Sigma L_{1i})(\Sigma L_{2k})}$$

For purposes of comparison with the variance of a specific rate, we will make additional simplifications to that we made previously and assume that all L_{1i} and L_{1k} are equal and that each is made up of $\frac{m}{n}$ periods of length L where m is the number of observation periods and n the number of persons observed. Then the covariance of p_1p_2 =

$$- \frac{P_1P_2\, n\, (\frac{m}{n}L)\,(\frac{m}{n}L)}{(mL)^2} = - \frac{P_1P_2}{n}$$

which is identical with the covariance of binomial proportions for the special case assumed here.

We now compare the absolute size of the covariance between two equal incidence rates with the variance of one of them.

$$\frac{\text{Covariance } (P_1P_2)}{\text{Variance } P_1} = \frac{P^2\,(mL)}{nP\,(1-PL)}$$

$$= \frac{P}{(1-PL)} \frac{mL}{n}$$

For P of .01, L of 2 years and each person contributing about 10 years experience, the ratio of covariance to variance is about 1 to 10. Recalling that we have artificially magnified the covariance by presuming that all persons have equal experience in both classifications, it is reasonable to ignore the covariance at least for the Framingham Study. For these data it is not sufficiently large in comparison to the variance to require its use when considering the standard errors of differences between rates.

IX

PANEL DISCUSSION: STATISTICS AND SURVEYS AS LEGAL EVIDENCE

Chairman, Eli S. Marks, Case Institute of Technology

Panel:

W. Edwards Deming, Consultant in Statistical Surveys
Arnold J. King, National Analysts, Inc.
James A. Bayton, National Analysts, Inc.

SESSION ON STATISTICAL AND SURVEY EVIDENCE IN THE COURTS
ANNUAL MEETING OF THE AMERICAN STATISTICAL ASSOCIATION

Eli S. Marks, Case Institute of Technology

This session is focused on the use of surveys and statistics as legal evidence. We are particularly interested in survey evidence and the problem of how one gets the courts to examine whether survey results are statistically correct rather than whether they are mathematically correct.

National Analysts has had some experience with this problem in the du Pont case where they did a survey regarding which Dr. Benjamin Tepping testified (before the United States District Court in Chicago). More recently, some of their staff appeared in a hearing before a Federal Trade Commission examiner with respect to a survey presented as evidence in a case involving the Borden Milk Company. In both these cases (as in others in which have involved survey evidence) the question of the admissibility of survey evidence has come up -- the critical point being, of course, that the idea of statistical data as distinct from individual testimony is a relatively novel concept in the courts. In both of the cases in which National Analysts was involved, the court and the examiner accepted their survey results as statistical results, over the objections of the government attorneys who persisted in trying to treat the survey data as individual reports subject to attack in terms of their individual validity.

The idea that data which are inaccurate when taken one at a time, may have quite adequate accuracy when considered as statistical aggregates is, of course, a fairly sophisticated concept and one which should (in my opinion) be expounded by the statistical profession as a whole rather than by individual statisticians. In organizing this session, I suggested that the discussion be slanted toward what can be done by the statistical profession as a whole to educate legal thinking on this matter.

Another phase of the discussion is the question of confidentiality and the assessment of the reliability of survey evidence. Dr. Eckler has outlined the recent Supreme Court decision which permits the government to subpoena the manufacturer's copy of reports submitted to the Census Bureau. The confidentiality of the Census Bureau's copy is, of course, legally protected. The decision in this case is at best a nuisance and at worst a negation of the primary purpose of the legal provision for confidentiality. That is, the purpose of confidentiality is to insure complete and truthful reporting to the Census Bureau by assuring respondents that their reports can never be used against them. Of course, the respondents can still protect themselves by not keeping copies of reports sent to the Census Bureau. This would, of course, mean difficulty in checking on incomplete or ambiguous entries. More important is the probability that respondents with anything to conceal will in the future supply statistical data only to the extent that compliance is legally required and legally enforceable!

The problem of confidentiality also came up in the cases in which National Analysts was involved. The opposition (government) attorneys have wanted access to individual reports -- their argument being that, if the court grants the admissibility of survey evidence (over their objection), they and the court should have an opportunity to assess the accuracy and validity of the evidence submitted. Adhering to the position that survey data constitutes "hearsay evidence" the government attorney in one case demanded the right to cross-examine the individual respondents in open court. National Analysts, of course, opposed this on the basis of the confidentiality of the information -- i.e., indicated that the data were secured under a promise of confidentiality and could (probably) not have been secured without this promise -- but I think we need, in the future, to go farther.

The principle of confidentiality must, of course, be upheld in its own right and as a matter of ethics and practical business. Beyond that, however, we need to get across to the courts and to the legal profession, that error in individual statistical reports does not demonstrate satisfactorily error in a statistic based on these reports. To do this we cannot "sit on our expert testimonies" and assert that the court must accept our statement that the data are accurate, without ever attempting to make an independent assessment of this fact. Making such an independent assessment is, of course, not a simple matter, but I do not agree with one opinion that has been expressed that the accuracy of survey data should be a matter of expert testimony with each side being allowed to present the opinions of their experts and may the best experts win! Certainly, we want to insist that the assessment of the accuracy of a survey is a matter of examination by experts, but we all know that the only way to begin to resolve a controversy with respect to the validity of a survey result (as distinct from its variance) is to repeat the survey on the same or a different sample -- i.e., demonstrate the reproducibility of the survey results.