COWLES COMMISSION DISCUSSION PAPER: ECONOMICS NO. 256

Some Notes on Aggregation and Cross-Section Studies

Discussion Remarks at the Conference on Research in National Income and Wealth, April 1 and 2, 1949

J. Marschak

April 1, morning

If data are not to be collected for their own sake, they are collected in order to permit conclusions about important questions. What is important depends on our set of values, that is, on our practical aims. We may deem it important to diminish the inequality of incomes, for reasons of envy, justice, or social peace: hence, our interest in income distribution taken by itself. Or it may seem important to us that year-to-year fluctuations of consumption be diminished, to stabilize employment: hence, our interest in combining data on income distribution with data on the effect of family income on family savings. And so on.

Ideally, we should like to know the multivariate distribution of households in which the variables are, say: current and past income of the family, its present and past assets, its various expenditures; total national income; local prices; occupations and ages of family members; geographical region and size of community, etc., etc. Suppose for a moment that this ideal were attained (which it never will be). Then it would be possible to predict the aggregate value for a certain variable when certain aspects of the multivariate distribution are changed in a pre-assigned way, while other aspects (certain "conditional distributions") remain unchanged. For example, suppose there is reason to assume that the distribution of family savings for given family size, income bracket, and fixed values of all other variables remains unchanged for twenty years ahead; while the distribution of income by family size, income brackets, etc. is changed in a way that is defined as the result of a proposed and debated fiscal or wage or farm price policy. Then our ideal multivariate distribution would help us to obtain

new weights for each of the cells, such as the cell called "Midwestern large-city tenant family of four people of whom two make $4,000 and $1,000 a year, and which owns a used car and $5,000 E-bonds"), and to derive the aggregate savings as a multiply-weighted sum of family savings. Analogous operations yield aggregate rent, food demand, etc.

This ideal will never be, and need not be attained. To be sure, an error is incurred whenever ignorance forces us to replace the ideal ten-variate (or is it hundred-variate?) distribution by a bi- or tri-variate one. That is, certain weighted sums are then replaced by unweighted or by wrongly weighted ones. When is this error serious? When it leads to a policy recommendation that fails its purpose. Our sense of what are and what are not the important policy alternatives must and does guide us when we allow ourselves to neglect the effect of certain variables and assume equal weights when they should be unequal; and when, on the other hand, we insist on deepening the analysis of those aspects of the multivariate distribution that we deem decisive in the determination of policies. In this, I see the gist of Dorothy Brady's paper and (though not always as outspokenly) of many other papers before us.

April 1, afternoon

In this age of microfilms and punching machines, it should be possible for any data-collecting agency to supply interested and competent people with copies of original schedules or punch cards, complete or selected at random. Then anyone could tabulate the data in whatever ways he wishes. This would be analogous to the biologists' custom of providing each other with specified strains of mice or bacteria. To circulate sets of 3,000 punch cards (the size of the Ann Arbor sample) would perhaps cost less than the combined travelling expenses of conferences where discussions have to be based, not on the full information contained in the data, but on tables in which part of the information is killed; and where

much of the time is spent in reviving the killed information, and in accusing or defending the murderers.

April 2, morning

The discrepancy in the estimates of the top-income share that are based, respectively, on field surveys and on tax returns, may be due to underrepresentation as well as underreporting. In both cases, there may be a bias (reluctance to admit an interviewer or to tell him the truth, if income is high and penalty nil); and there may be a sampling error. One would like to learn more as to how the biases and sampling errors are minimized when a survey is designed, and how they are estimated when the results are computed. In a sample of 3,000, the probability is about 25% that 18 or less (instead of 30) people above the top percentile of the population are interviewed; and that, consequently, when such a sample is "blown up," 9% or less (instead of Kuznets' 15%) of national income is estimated to go to the top 1%. When applied to savings, this random difference is much larger still. Furthermore, the probability is still about 6% that this error is repeated in two subsequent field surveys. While I do not insist that the discrepancy is solely due to a sampling error, and not to a systematic bias and to underrepresentation rather than underreporting, I urge that the distribution of all errors be estimated and freely discussed by those who publish results of field surveys.