Two-sample cross tabulation

Tomoki Fujii, Singapore Management University
Roy van der Weide, World Bank

"Cross-tabulation is a useful descriptive tool that is routinely used in economics, sociology, political science and other disciplines to describe the relationship between two variables of interest for a given unit of observation. It is easy to produce and helpful for understanding the relationship between two discrete variables of interest. Typical units of observation are individuals, households, firms, agricultural plots, etc. For ease of exposition, we will refer to observations as ""households."" The only requirement for computing a cross-tabulation is that these two variables are available for the same households. However, variables of interest are often divided over different samples that cannot be linked with certainty at the household level. Surveys tend to be specialized and are undertaken by different parties. It is often not in the interest of those commissioning the survey to collect data beyond that what is relevant for their program or project. Even if funding and coordination are not an issue, grouping multiple specialized surveys into one so that all data are collected for the same sample of households will be demanding both on those interviewed and interviewed. This may compromise the quality of the data collected.

When no one sample contains both variables of interest, a popular approach is to pick one sample and impute the variable that is missing or use a proxy for the variable of interest, such as an asset index, which does not have a straightforward interpretation. It is assumed that the samples have a set of variables in common in addition to the variables that are unique to each sample. The shared variables then serve as instruments in the model used for imputing the missing variable. This means that exact observations (for variable one) will be cross tabulated with imputed values (for variable two). We will refer to this as a two-sample cross-tabulation.

Two-sample cross-tabulations are subject to two types of error: imputation error and sampling error. The added imputation error will add to the standard error of the estimates. It will generally also introduce a bias, as the imputation error is likely to be correlated with the variable with which it is cross-tabulated. The dual error structure is often ignored in empirical applications, leading to an overestimation of the statistical precision of cross-tabulations that feature imputed values.

This paper puts forward a bias-corrected estimator and derives its asymptotic distribution based on the prior regarding the correlation in unobservable error terms. The asymptotic distribution takes into account both the imputation- and the sampling error. These analytic results enable the user to compute standard errors without having to resort to bootstrap simulations. To ensure exact standard errors for any sample size and sampling design, we also offer a bootstrap procedure. Monte Carlo simulations suggest that the asymptotic- and bootstrap standard errors are equally accurate also for small sample sizes; both match the true standard errors. Having analytic standard errors that are both easy to compute and accurate makes two-sample cross-tabulation a user-friendly tool for a wide audience of applied users.

The contexts in which two-sample cross-tabulations can be applied are wide-ranging. Two prominent examples of specialized surveys are the household budget survey (HBS) and the demographic health survey (DHS). The HBS collects detailed household expenditure data that is

commonly used to determine a household's poverty status. The DHS collects detailed health and healthcare data that includes anthropometric indicators, whether the individual is HIV positive, and usage of health services. While basic demographics, education, asset ownership, and dwelling unit characteristics are typically available in both types of surveys, the detailed expenditure and health data are unique to the corresponding surveys. A cross-tabulation of health with poverty would then require a combination of the two surveys. Specifically, one may wish to examine the relationship between socio-economic status and child malnutrition, child health care, and usage of health services more generally. However, asset index, and not consumption measure, is typically used as a measure of socio-economic status, simply because the latter is unavailable in the DHS. Our approach resolves this issue.

Another useful example is the tracking of welfare over time offers another example. For example, one may wish to produce the poverty transition matrix, which tells the probability of falling into [getting out of] poverty in the next period, given that the individual is non-poor [poor]. This matrix is conventionally only available if one has panel data. The two-sample analog is obtained by combining repeated cross-sections to create a so-called synthetic panel. Consider a table that shows the percentage of households falling into and out of poverty which helps address the question of whether poverty is of a chronic or a transient nature. Similarly, one may study the transitions in terms of food security, malnutrition, employment status, location of residence (urban vs rural), etc. Note that transition matrices can also be estimated from repeated cross-sections without the use of survey-to-survey imputation. Under some rigid identifying assumptions, one could estimate the transition probabilities directly by means of maximum likelihood estimation (Moffitt, 1993), but it is difficult to tell how these probabilities change when these assumptions are violated. Our approach overcomes this issue by parameterizing our results with the correlation in unobservable error terms, which is intuitive and enables us to check the robustness of our results.

We provide numerical simulation results as well as empirical estimation results using household survey data to illustrate our method. In particular, we show that the bias correction helps to arrive at reliable estimates of the relationship between subjective poverty and consumption poverty. To our knowledge, this is the first paper to provide bias-corrected estimators for two-sample cross-tabulation estimators and analytic standard errors for two-sample cross-tabulation. Our contribution also has a practical value, because the analytic standard errors presented in this paper are easily implemented without the use of bootstrapping."