Sampling and Data Analysis: Theory and Applications to Agriculture

by

Jean-Paul Chavas

University of Wisconsin, Madison

Preliminary Draft

Sampling and Data Analysis: Theory and Applications to Agriculture

1. Introduction

Data collection is at the heart of knowledge creation and its use in private decision making, public policy and applied research. Yet, knowing what data to collect and how to analyze it can be difficult. Statistical methods have helped guide the process of data collection and analysis. They have provided valuable tools that help us become more effective in improving our understanding of the economy. This is particularly important in the social sciences as economic systems involve complex interactions between private agents, public institutions and markets. It means a strong need to collect and process information about the evolving status of economic systems to inform both private and public decision makers. And the availability of economic data is crucial to academic researchers studying the functioning of market economies and the efficiency of managerial and policy choices.

An integrated analysis of data availability, choices, survey cost, and statistical methods can improve the flexibility, precision and usefulness of data collection and analysis. This paper presents a brief overview of these issues. First, it reviews what statistical theory offers as guidance in the process of collecting and analyzing data. This covers the evaluation of survey design and summary measures related to both attributes and choices made by particular decision makers. It also covers the econometric analysis of decision rules made by individuals as part of the functioning of the economy, and the evaluation of welfare outcomes. Second, the paper examines the optimality of data collection and analysis. Such optimality always depends on the information we are looking for. And in the social sciences, data collection is also constrained by the fact that investigators typically have incomplete experimental control. Third, implications for the agricultural sector are briefly presented. This includes a discussion of the definition of a “farm” for data collection purpose, both from a statistical viewpoint and an economic viewpoint. Arguments are presented stressing the role of microeconomic dynamics, and the need for better panel data to help us assess the role of managerial ability and its effects on economic adjustments to changing market conditions and technology.

2. Sampling about attributes and choices

This section sets up the notation for the rest of the paper. We consider an economy involving N individuals (e.g. farmers), where each individual faces a vector of attributes z and chooses a vector y from a feasible set Y, where $y \in Y$. Let $\mathcal{N} = \{1, \ldots, N\}$, where $\mathcal{N}$ is the "population" of interest. The attributes z reflect the economic environment as well as the personal characteristics of each decision maker. At the time when the decisions y are made, let $z = (z_a, z_{b,}, z_c)$, where z_a is observed by both the decision maker and the investigator, z_b is observed by the decision maker but not the investigator, and z_c remains unobserved for the decision maker and the investigator.[1] Let F(z) be the distribution function of z in the population. Denote the associated marginal distribution of z_a by $F(z_a)$ and the conditional distribution of z_c given (z_a, z_b) by $F(z_c \mid z_a, z_b)$. Assume that each decision maker has preferences represented by a von Neumann-Morgenstern utility function U(y, z). Under the expected utility model, the choice of y made by an individual with attributes (z_a, z_b) is given by[2]

$$V(z_a, z_b) = \text{Max}_y \; \{ \int_{-\infty}^{\infty} U(y, z) \, dF(z_c \mid z_a, z_b) \colon y \in Y \}, \qquad (1)$$

which has for solution $y^*(z_a, z_b) \in \text{argmax}_y \; \{ \int_{-\infty}^{\infty} U(z, y) \, dF(z_c \mid z_a, z_b) \colon y \in Y \}$. An analysis of the population involves the variables $z = (z_a, z_b, z_c)$ and $y^*(z_a, z_b)$ under the distribution F(z). As noted above, the investigator observes z_a and y^*, but does not obverse (z_b, z_c). The question is: how can we collect data from the population that would provide useful information?

Answering this question depends on what kind of information we want. There are there pieces

[1] This implicitly assumes that the individual is better informed than the investigator (as reflected by the vector z_b).

[2] It means that $F(z_c \mid z_a, z_b)$ is interpreted as the probability distribution of z_c as seen by an individual with attributes (z_a, z_b).

of information that may be of interest. First, we may want to know what is happening in the aggregate population. This typically involves evaluating population means for z_a and y^*:

$$E_a(z_a) = \int_{-\infty}^{\infty} z_a \, dF(z_a), \tag{2a}$$

and

$$E_{ab}(y^*) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^*(z_a, z_b) \, dF(z_a, z_b), \tag{2b}$$

where E_a and E_{ab} are expectation operators based on the (marginal) distributions $F(z_a)$ and $F(z_a, z_b)$, respectively. And when z_a and y are quantities or monetary values, we may want to know the aggregate population totals $[N\, E_a(z_a)]$ and $[N\, E_{ab}(y^*)]$.

Second, we may want to assess the factors affecting the choices y^*. This typically involves investigating the "regression lines":

$$E_{b|a}(y^* \mid z_a) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^*(z_a, z_b) \, dF(z_b \mid z_a), \tag{3}$$

where $E_{b|a}$ is the expectation operator based on the conditional distribution $F(z_b \mid z_a)$.

Third, we may want to investigate the welfare of individuals in the population. This is given by the function $V(z_a, z_b)$ in equation (1). When $V(z_a, z_b)$ is observable (e.g., the case of profit), it can be used to evaluate average welfare per capita

$$E_{ab}(V(z_a, z_b)), \tag{4}$$

or aggregate welfare $[N\, E_{ab}(V(z_a, z_b))]$.

Data collection starts with two crucial steps: 1/ the definition of population; and 2/ the identification of a sampling frame. Because observing the whole population is costly, the goal is often more modest: finding a representative sample of the population (or a subset of the population). The choice of the population (of its subset) depends on what information we are looking for. So does the choice of the samplings scheme.

There are two broad categories of sampling schemes. The first category is exogenous sampling, when the investigator selects decision makers and observe their choices. Using our notation, it means selecting individuals before we know what decisions y they make. Then, data are collected on the attributes z_a and the decisions y^* made by each selected individual in the sample. The second category is choice-based sampling, when the investigator selects alternatives and observe the decision makers choosing them. In this case, individuals are selected depending on their observed decision $y^*(z_a, z_b)$. Again, data are collected on the personal attributes z_a and the decisions $y^*(z_a, z_b)$ made by each selected individual in the sample.

3. Data analysis

Data collection and data analysis are part of the process of obtaining information about the population of interest.

3.1. Estimating population means or totals

Consider a sample of n individuals taken from the population of N individuals, with $n \leq N$. When $n < N$, a sampling scheme involves choosing a subset of individuals, the i-th individual being selected out of the population $\mathcal{N} = \{1, \ldots, N\}$ with probability $p_i \in (0, 1]$, $i \in \mathcal{n}$, where $\mathcal{n} = \{1, \ldots, n\} \subset \mathcal{N}$. We may want to use this information to estimate the populations means $E_a(z_a)$ in (2a) and $E_{ab}(y^*)$ in (2b).

First, consider the case of a random sample, where each individual is chosen independently and with the same probability: $p_i = p$, $i \in \mathcal{n}$. This is the simplest form of exogenous sampling. It has some desirable characteristics. Let (y_i, z_i) be the values taken by (y, z) for the i-th individual in the sample, $i \in \mathcal{n}$. The sample mean of z_a, $\sum_{i \in \mathcal{n}} z_{ai}/n$, is an unbiased and consistent estimator of $E_a(z_a)$ in (2a); and the sample mean of y^*, $\sum_{i \in \mathcal{n}} y_i^*/n$, is an unbiased and consistent estimator of $E_{ab}(y^*)$ in (2b). These sample estimates have variances that are inversely proportional to n. And they are asymptotically normal (from the central limit theorem). This provides a convenient way to obtain useful information on population means or population totals, with an accuracy that can be controlled through the choice of the sample size n.

In general, the choice of the sample size n can be analyzed as an economic decision. To illustrate, consider the case of a random sample $\{z_{a1}, .., z_{an}\}$, where the variance of z_a is $\mathrm{Var}(z_a) = \sigma_a^2 > 0$. Then, the variance of the sample mean $(\sum_{i\in n} z_{ai}/n)$ is σ_a^2/n. Let the cost of sampling be c(n) and the (gross) benefit of the sample information be $b((\sigma_a^2/n)^{-1})$, with $c' = \partial c(n)/\partial n \geq 0$ and $b' = \partial b(k)/\partial k \geq 0$. The optimal sample size n is the one that maximizes net benefit: $n^* \in \mathrm{argmax}_n \{b((\sigma_a^2/n)^{-1}) - c(n)\}$. Assuming an interior solution, the optimal sample size n^* satisfies the first-order condition:

$$b'/\sigma_a^2 = c'. \tag{5}$$

Equation (5) simply states that the optimal sample size n^* is obtained when marginal cost is equal to marginal benefit. In situations where marginal cost is zero ($c' = 0$), this would imply that the optimal sample size n^* would correspond to the point where marginal benefit is also zero, with $b' = 0$. Alternatively, when b' is decreasing in n, the optimal sample size n^* would be reduced when the marginal cost of sampling is positive, $c' > 0$.

Second, consider the case where the p_i's vary across individuals in the sample, corresponding to a non-random sample. When the p_i's are all positive, the sampling weight for the i-th individual in the sample is $w_i = 1/p_i$, $i \in n$. Non-random samples can be useful. One example is the case where the investigator is interested in particular subsets of the population. Then, a stratified sample can be used, where the probability p_i of being selected increases in the subsets of interest (while treating the individuals within each subset as in random sampling). This can yield better information on the individuals of interest. Then, the appropriate sample means are the weighted mean $\sum_{i\in n} (w_i\, z_{ai})/(\sum_{i\in n}$

w_i) which provides a consistent estimator of $E_a(z_a)$ in (2a), and the weighted mean $\sum_{i\in n} (w_i\, y_i^*)/(\sum_{i\in n} w_i)$ which is a consistent estimator of $E_{ab}(y^*)$ in (2b). As long as the sampling scheme remains exogenous, these nice properties continue to hold.[3]

Third, consider the case of non-exogenous sampling. Then, the unbiasedness and consistency properties may not hold. This can arise from self-selection by non-respondents. Indeed, if some targeted individuals refuse to participate and their decision to self-select out of the sample is correlated with the attribute z_a and/or the choices y^*, then the sample may no longer be representative of the population. In this case, the sample means would be biased estimates of the corresponding population means.

And similar problems can arise under choice-based sampling. Consider the case where the investigator targets individuals who make decisions that belong to a particular choice set S. Only individuals who make decisions that satisfy $y^*(z_{ai}, z_{bi}) \in S$ can become part of the sample, implying that the sample is selected from the subset $I = \{i: y^*(z_{ai}, z_{bi}) \in S, i \in N\}$. What are the properties of the associated sample means for z_a and y^*? There is one situation where the analysis remains simple. This occurs when two conditions hold: 1) we have random sampling within the subset *I;* and 2) $y^*(z_{ai}, z_{bi}) \notin S$ implies that $y^*(z_{ai}, z_{bi}) = 0$ for all $i \in N$. Then, the sample means $\sum_{i\in I} y_i^*/n$ provide an unbiased an consistent estimate of $E_{ab}(y^*)$. Otherwise, arithmetic sample means would give a biased estimate of the associated population means. Let the probability that the i-th individual gets selected in the sample be $Pr_i = Prob[y^*(z_{ai}, z_{bi}) \in S]$. When the Pr_i's are all positive and vary across individuals and letting $W_i = 1/Pr_i$, consistent estimates of the population means can be obtained using weighted means: the weighted mean $\sum_{i\in I} (W_i\, z_{ai})/(\sum_{i\in I} W_i)$ gives a consistent estimate of $E_a(z_a)$ in (2a), and the weighted mean $\sum_{i\in I} (W_i\, y_i^*)/(\sum_{i\in I} W_i)$ provides a consistent estimate of $E_{ab}(y^*)$ in (2b). Alternatively, when $Pr_i = 0$

[3] Another example is the case of cluster sampling which involves selecting individuals in groups. This is typically motivated by the convenience of data collection (e.g., when the group is defined by its location). But it raises the issue of possible intra-cluster correlation across observations that needs to be taken in consideration in statistical analysis.

for some i ∈ *N*, sample information is subject to selection bias as a part of the population remains unobserved. The issue of selection bias is further discussed below.

3.2. Estimating regression lines

In general, $y^*(z_a, z_b)$ in equation (1) is a decision rule mapping particular situation represented by (z_a, z_b) into a choice y^*. The variables (z_a, z_b) can be individual specific: by representing differences in individual attributes, they would then characterize individual heterogeneity. Or they can be common to some or all individuals, thus representing the economic environment of decision makers. In this case, the decision rules $y^*(z_a, z_b)$ inform us how changes in economic conditions affect behavior. Economists are often interested in both individual heterogeneity and behavioral response to economic changes. Since the attributes z_b are assumed to be unobserved by the investigator, it means that we cannot obtain direct information on their effects. Thus, the analysis typically focuses on the effects of the observable attributes z_a. This is where the regression lines $E_{b|a}(y^* \mid z_a)$ in (3) comes in. The conditional expectation $E_{b|a}(y^* \mid z_a)$ in (3) evaluates the expected value of the choices y^* given attributes z_a. It provides information on how decisions can change under alternative situations (represented by z_a).

How can we evaluate $E_{b|a}(y^* \mid z_a)$ using sample data? A common way to proceed is to assume a parametric structure for this conditional expectation: $E_{b|a}(y^* \mid z_a) = f(z_a, \beta)$, where $\beta \in R^k$ is a vector of k parameters. The "regression lines" are the set of values taken by $f(z_a, \beta)$ as the variables z_a change. Then, the statistical analysis consists in using sample data to estimate β. This generates the following model

$$y_i^* = f(z_{ai}, \beta) + e_i, \qquad (6a)$$

i = 1, …, n, or after stacking all observations,

$$\mathbf{y} = \mathbf{f}(Z_a, \beta) + \mathbf{e}, \qquad (6b)$$

where $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{f}(Z_a, \beta) = \begin{bmatrix} f(z_{a1}, \beta) \\ \vdots \\ f(z_{an}, \beta) \end{bmatrix}$, and $\mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$. Equation (6) is a standard regression model where **e** is an "error term" distributed with mean zero (e.g., Greene 2007).

Assume that $\mathbf{f}(Z_a, \beta)$ is differentiable in β, with $\Delta_\beta\mathbf{f}(Z_a, \beta) \equiv \partial\mathbf{f}(Z_a, \beta)/\partial\beta$ denoting the matrix of derivatives of $\mathbf{f}(Z_a, \beta)$ with respect to β. Below, we consider the following conditions associated with model (6):

Condition A1: $\text{Rank}[\Delta_\beta\mathbf{f}(Z_a, \beta)] = k$.

Condition A2 (homoscedasticity): $E(e_i\, e_j) = \begin{Bmatrix} \sigma^2 \\ 0 \end{Bmatrix}$ when $\begin{Bmatrix} i = j \\ i \neq j \end{Bmatrix}$.

Condition A3 (orthogonality): $E[\Delta_\beta\mathbf{f}(Z_a, \beta)\, \mathbf{e}] = 0$.[4]

Condition A1 requires that $n \geq k$, i.e. that there is enough sample information to estimate k parameters. We assume that A1 is satisfied throughout the paper. The choice of an econometric estimator for β in (6) depends in large part on conditions A2 and A3. First, consider the case where both A2 and A3 holds. Then, the ordinary least-squares (OLS) estimator of β in (6) is consistent, efficient, and asymptotically normal (from the central limit theorem).

Second, consider the case where condition A3 holds, but A2 does not. Then the OLS estimator of β in (6) is consistent but inefficient. If OLS is used, any hypothesis testing should be based on "robust" standard errors, i.e. standard errors that are corrected for the inefficiency of the OLS estimator (e.g., Greene 2007, p. 462). Alternatively, when the homoscedasticity assumption A2 does not hold, β can be estimated by weighted least-squares. With an appropriate choice of weights (e.g., weights being inversely proportional to the standard deviation of e_i under heteroscedasticity), weighted least squares would provide an estimate of β that is consistent and efficient.

Third, consider the case where condition A3 does not hold. Then, the OLS estimation of (6)

[4] To be conformable, the matrix $\Delta_\beta\mathbf{f}(Z_a, \beta)$ is defined to have k columns and a number of rows equal to the dimension of **e**.

would provide an estimate of β that is biased and inconsistent. This is a case of "endogeneity bias" where the unobserved determinants of both y^* and z_a are correlated. The solution is to rely on an instrumental variable estimator (e.g., the generalized method of moments or GMM estimator; see Hansen 1982), assuming that proper instruments can be found to deal with the endogeneity problem.

4. Sample design

The above arguments related to the estimation of β in (6) apply under general conditions. In particular, they apply whether or not the data used in (6) come from a random sample. But are there any linkages between the sampling scheme and the estimation of the parameters β in (6)? This issue has been addressed in the literature on optimal experiment design (e.g., Atkinson and Donev 1992).

4.1. Optimal experimental design

In the context of model (6), the "optimal experiment" is the one generating data that provide the most accurate information about the parameters β. Since the parameters β are specific to the model in (6), this means that the optimality of a sampling scheme depends on the situation being analyzed.

Note that there are situations where a random sample corresponds to a "poor" experimental design. This can happen under two scenarios: 1) when the population exhibits very little variation in some of the variables in z_a; and 2) when some of the variables in z_a tend to move together in the population. In the first scenario, a lack of variation in some explanatory variables makes it difficult to estimate the regression parameter(s) measuring their marginal effect on y^*. In the second scenario, random sampling would generate a collinearity problem that generates imprecise estimate of some of the parameters. In either case, the "optimal experiment" would be to move away from random sampling toward a sampling scheme that increases the range of observations for selected variables (in scenario 1) and/or to "oversample" in regions where the collinearity problem can be reduced (in scenario 2). But this assumes that we have (at least some) experimental control.

4.2. Sample design without experimental control

A major issue in the social sciences (compared to the so-called "hard sciences") is the fact that human behavior is complex and is typically not subject to direct control by the investigator. In the absence of experimental control, we often do not have the luxury of designing an "optimal experiment". Especially when the number of relevant variables in z_a is large, collinearity can make it quite difficult to obtain reliable estimate of the parameters β in (6). This has stimulated economic inquiries in three directions: 1) conducting "real world" experiments (as experimental economics has become widely used in the empirical investigation of human behavior; e.g., Smith 2008); 2) looking for "natural experiments" that come a little closer to an "optimal experiment" (e.g., DiNardo 2008); 3) designing randomized field experiments that can help in the identification and estimation of key parameters in (6) (e.g., Duflo et al. 2007).

The above discussion makes it clear that random sampling is neither necessary nor sufficient to obtain good estimate of the parameters β in (6). A corollary is that, under stratified sampling, there is not direct linkage between the sampling weights w_i and the regression weights used in the estimation of (6). The former weights are relevant in the estimation of the population mean $E_{ab}(y^*)$, while the latter may be used in the estimation of the parameters β characterizing the condition mean of $(y^* \mid z_a)$ in (6). The optimal regression weights in (6) depend on the variance of **e**, i.e. on the conditional variance of y^* given z_a. These weights can be used to improve the efficiency of the parameter estimate when condition A2 does not hold (e.g., under heteroscedasticity where the variance of e_i is not constant).[5] Note that there can be situations where condition A2 holds under stratified sampling, in which case there would be no need to use regression weights in the estimation of β in (6). Alternatively, there can be situations where condition A2 does <u>not</u> hold under random sampling, in which case efficiency

[5] As noted above, if OLS is used when A2 does not hold, then any hypothesis testing related to β should rely on "robust" standard errors.

considerations would suggest the need to use regression weights in the estimation of β in (6).

This shows that the estimation of population means in (2) and the estimation of the regression lines in (3) or (6) raise different statistical issues. The former relies on estimating unconditional means, while the latter involves estimating the conditional mean $E_{b|a}(y^* \mid z_a)$. Yet, the two are related. Indeed, they satisfy $E(y^*) = E_a[E_{b|a}(y^* \mid z_a)]$. It means that it is always possible to go from (conditional) regression results to the assessment of the unconditional population mean for y^*. And under stratified sampling, note that the sampling weights always play a role in using sample information to estimate population means or population totals.

What about the case of non-exogenous sampling? A new problem can arise under choice-based sampling. Indeed, when the unobservable factors z_b affect both the prospects for an individual to be selected in the sample and his/her decision y^*, then the OLS estimation of β in (6) can give inconsistent parameter estimate due to "selectivity bias". This selectivity problem has been identified and discussed by Heckman (1979). Following Heckman (1979) and others, econometric methods have been proposed to obtain consistent estimate of the parameters β in (6) under selectivity and choice-based sampling.

5. Evaluating Economic Welfare

Economists are often interested in welfare evaluation. This can be done by evaluating the function $V(z_a, z_b)$ in equation (1). Since z_b is not observable, the best we can do is to evaluate the distribution $V(z_a, z_b)$ in the sample and/or the population, treating z_b as a random variable. Summary measures of this distribution are given by its conditional mean $E_{b|a}(V(z_a, z_b))$ (conditional on z_a) and its unconditional mean $E_{ab}(V(z_a, z_b))$ in (4). These can be interpreted as average welfare measures per capita. Alternatively, one can evaluate the associated aggregate welfare $[N\ E_{ab}(V(z_a, z_b))]$.

Evaluating (1) or (4) from sample information is relatively easy when the variable V is observed (e.g., profit). Then, the discussion presented in sections 3 and 4 also applies to empirical welfare analysis. But individual welfare is often not directly observable. Prevalent examples include consumer preferences and risk preferences. In such situations, the evaluation of V must be indirect. The analysis then must rely on "revealed preferences" and the linkages between observed behavior $y^*(z_a, z_b)$ and $V(z_a, z_b)$ given in (1). Such linkages can be used to recover information about the underlying individual preferences. Good progress has been achieved making these linkages empirically tractable. Notable examples include the Almost Ideal Demand System (AIDS) of Deaton and Muellbauer (1980), and the Quadratic Almost Ideal Demand System (QAIDS) of Banks et al. (1997) to recover information about the preferences and welfare of consumers. Other examples include the analysis of risk preferences and the empirical assessment of the cost of risk (e.g., Chavas and Holt 1996) and the investigation of discrete choices (e,g, McFadden and Train 2000). Once the linkages between observed behavior and preferences are made, the analysis presented in sections 3 and 4 can be readily extended to support an empirical evaluation of individual and aggregate welfare.

6. Applications to agriculture

Much interest has focused on evaluating the performance of agriculture and its ability to feed a growing world population. The analysis presented in sections 3-5 would apply broadly to the assessment of the agricultural and food sector, including production and environmental management, productivity and sustainability, the structure of agriculture, food pricing and marketing, and consumer behavior and welfare.

In this context, an important empirical issue is to estimate the economic contribution of agriculture to aggregate welfare. When aggregate welfare is measured by the Gross Domestic Product (GDP), this requires assessing the monetary value of agricultural commodities produced by the farm sector. In turns, this requires a definition of "a farm". There is no clear consensus about the definition of a farm. Since 1850, the US Census definition of a farm has changed nine times. For statistical purposes, the following definition has been used by USDA, the Office of Management and Budget and the US Census Bureau since 1975: “A farm is any place from which $1000 or more of agricultural products were sold or normally would have been sold during the year under investigation”.

Why use $1000 as a threshold measure? This is an attempt to capture the fact that a large part of agricultural production comes from larger farms. On the one hand, it seems reasonable to argue that not every household that has a “small garden” should be treated as a farm. On the other hand, identifying a proper dividing line between a “large garden” and a “small farm” is not obvious. Note that such an issue is particularly relevant when we consider the case of “hobby farms” and of “urban agriculture”. Both hobby farms and urban agriculture can produce agricultural outputs that are sold in the market place. When their agricultural output is positive but worth less than $1000, why should their contribution to agricultural production be neglected?

To shed some light on these questions, we present two sets of arguments: 1) there is no statistical basis for choosing a $1000 minimum threshold in the definition of a farm; but 2) there is an economic basis to choose such a threshold. To see that, start with the main objective: to document the economic contribution of agriculture to GDP. Let $y^*(z_a, z_b) \geq 0$ measure the value of agricultural outputs produced and sold by a household with attributes (z_a, z_b) during a particular year in a given region. Then, the aggregate value of agricultural production is $[N\ E_{ab}(y^*)]$ where $E_{ab}(y^*)$ is given in (2b) and N is the total number of households in the region. Denote the distribution of y^* by $G(y) = \text{Prob}[y^*(z_a, z_b) \leq y]$.

Consider the case of a census. In general, we have

$$N\ E_{ab}(y^*) = N \int_0^\infty y\, dG(y) \geq N \int_s^\infty y\, dG(y), \tag{7}$$

where $s \geq 0$ is a minimum farm-level threshold used in the measurement of the value of agricultural outputs (e.g., s = $1000). The inequality in equation (7) reflects the fact that the agricultural outputs produced by households below the threshold s are not counted. In other words, when the distribution of output value is truncated at s, the associated truncation bias is:

$$B(s) = N \int_{0}^{\infty} y\, dG(y) - N \int_{s}^{\infty} y\, dG(y)$$

$$= N \int_{0}^{s} y\, dG(y) \geq 0. \qquad (8)$$

The truncation bias B(s) in (8) measures the part of aggregate production that is neglected in the census. Can this truncation bias be avoided? Yes. Choosing a threshold s = 0 would imply B(s) = 0 from (8). It simply means that, from a statistical viewpoint, avoiding any bias in the measurement of the aggregate value of agricultural outputs implies the need to count all households that produce agricultural products. If so, why would we want to choose a positive farm-level threshold (s > 0) in the economic evaluation of agricultural activities? Why would we want to tolerate a positive bias B(s) in the measurement of the aggregate value of agricultural production?

These questions can be answered if we consider that information collection is costly. Indeed, for a given minimum threshold s, the number of households remaining in the population (after truncation) is

$$n = N \int_{s}^{\infty} dG(y) \leq N. \qquad (9)$$

Equation (9) shows that increasing the threshold s reduces the number of households to survey in a census. Without a loss of generality, one can always ignore the households that produce no agricultural outputs (where $y^* = 0$). But choosing s > 0 allows a further reduction in the number of households that need to be surveyed. Assume that the cost of surveying n households is C(n), and that the (gross) social benefit of collecting information on the economic value of agriculture is T(-B(s)), with with $C' = \partial C(n)/\partial n \geq 0$ and $T' = \partial T(k)/\partial k \geq 0$. Then, the choice of a threshold s can be given an economic interpretation. The optimal threshold s is the one that maximizes net benefit:

$$s^* \in \text{argmax}_s \{T(-B) - C\} = \text{argmax}_s \{T[-N \int_0^s y\, dG(y)] - C[N \int_s^\infty dG(y)]\},$$

using (8) and (9). Assuming differentiability and an interior solution, the optimal threshold s^* satisfies the first-order condition:

$$T'\, s^* = C'. \tag{10}$$

Equation (10) simply states that the optimal threshold s^* is obtained when marginal cost is equal to marginal benefit. In situations where marginal cost is zero ($C' = 0$), this would imply that the optimal threshold S^* would also be zero: $s^* = 0$ .Thus, if collecting information is costless, it would be optimal to survey <u>all</u> households. But this result no longer applies when information is costly. Indeed, when $C' > 0$ and $T' > 0$, equation (10) implies that $s^* > 0$. Then, the optimal threshold s^* is positive, reflecting the fact that obtaining information from very small farms may not be worth the cost. In addition, equation (10) provides some useful insights on the factors affecting s^*. For example, if T' is decreasing, any increase (decrease) in C' would be associated with an increase (decrease) in s^*.

The above analysis focused on census data where all farms satisfying $y^* \geq s$ are surveyed. Note that this is a form of choice-based sampling discussed in section 3, where $I = \{i\colon y^*(z_{ai}, z_{bi}) \geq s,\ i \in N\}$. In this context, the bias identified in equation (8) is just an example of selection bias. Our discussion indicates that the definition of a farm (according to the minimum threshold s) involves an economic trade off between the cost of collecting information and its value. While a narrow statistical focus may suggest avoiding any bias in measurements, a broader economic approach (as shown in equation (10)) indicates that it can be optimal to tolerate some "small bias" when information cost is taken into considerations. Finally, while these arguments where developed in the context of a census, note that they can be extended to other sampling schemes where the sample is a subset of $I = \{i\colon y^*(z_{ai}, z_{bi}) \geq s,\ i \in N\}$.

7. Dynamics

The above discussion was presented at a given point in time. This is fine when there is no significant changes over time. But our world changes constantly, and a significant issue for economists is to understand the factors that drive these changes. This suggests refining our inquiry to consider dynamics. This can be done by introducing time explicitly in the analysis.

For that purpose, let $z_a = (z_{a0}, z_{aL}, y_L)$ and $z_b = (z_{b0}, z_{bL})$, where (z_{a0}, z_{b0}) are the values taken by the attributes (z_a, z_b) at the current time, (z_{aL}, z_{bL}) are the values taken by the attributes (z_a, z_b) in previous periods, and y_L denotes the lagged values of y. This captures dynamics in three possible ways: through lagged choices y_L; through the lagged attributes observed by both the decision maker and the investigator, z_{aL}; and through the lagged attributes observed only by the decision maker (and not the investigator), z_{bL}.

First, introducing lagged values $(z_{aL}, y_{L,} z_{bL})$ in the analysis does not affect the empirical assessment of population means (2a)-(2b) or their associated population totals. For example, under exogenous random sampling, sample means and sample totals would still provide unbiased and consistent estimates of the corresponding population means and totals (as discussed in section 3). But economic dynamics can be captured by the evaluation of the regression lines given in equation (3) or (6a)-(6b). Indeed, given $E_{b|a}(y^* \mid z_{a0}, z_{aL}, y_L) = f(z_{a0}, z_{aL}, y_L, \beta)$, equation (6) can now be written as

$$y^* = f(z_{a0}, z_{aL}, y_L, \beta) + e, \tag{11}$$

where $e \equiv y^*(z_{a0}, z_{aL}, y_L, z_{b0}, z_{bL}) - E_{b|a}(y^* \mid z_{a0}, z_{aL}, y_L)$. When $f(z_{a0}, z_{aL}, y_L, \beta)$ is differentiable in y_L, it follows that $\partial f(z_{a0}, z_{aL}, y_L, \beta)/\partial y_L$ captures the dynamics of microeconomic decisions under situation (z_{a0}, z_{aL}, y_L). There is much interest in such dynamics. In the context of the agricultural and food sector, three topics are particularly relevant: 1) technology adoption ; 2) economic adjustments to changing market conditions; and 3) the dynamics of obesity. All three topics require a good understanding of individual dynamics. Who are the early adopters versus late adopters of a new technology (e.g., Griliches 1957)? Who is better at adjusting to fluctuating markets (e.g., Huffman 1977)? And why is it that some individuals become obese and others do not? In each case, one wants to assess the role of managerial ability and how it varies across individuals. Addressing these issues require access to panel data that follow individuals over time.

Fortunately, the econometrics of panel data is reasonably well developed (e.g., Mundlak 1978; Hausman and Taylor 1981; Wooldridge 2002). The estimation of equation (11) using panel data raises additional challenges. As noted above, dynamics shows up in (11) in three ways: through y_L, through z_{aL}, and through z_{bL}. Note that, being unobserved, z_b and z_{bL} enter equation (11) only through the error term e, thus potentially creating serial correlation. In this case, condition A3 would fail to hold due to the correlation between y_L and e, thus exposing the estimation of β to endogeneity bias. Dealing with this issue includes using "fixed effects" and/or instrumental variable estimators (e.g., Mundlak 1978; Hausman and Taylor 1981; Wooldridge 2002). In addition, individual heterogeneity in the unobservables z_b implies that condition A2 would likely fail to hold. This suggests that the estimation of equation (11) should rely on robust standard errors and/or make use of appropriate regression weights to capture efficiency gains (e.g., as in "random effects" models).

Unfortunately, panel data are rather rare. In US agriculture, the best panel data available are the census data, collected every five years by the National Agricultural Statistics Service (NASS) at the US Department of Agriculture (USDA). Yet, having access to panel data can be quite valuable. A good illustration is given by Huffman (1977) who uses census data to document the role of human capital in the microeconomic dynamics of US agriculture. At this point, some important issues remain poorly understood in large part due to a lack of annual panel data. This includes the role of managerial skills in technology adoption, the role of human capital in market dynamics, and the role of education in the current obesity epidemic.

References

Atkinson A.C. and A.N. Donev (1992). *Optimum Experimental Designs*. Oxford University Press, Oxford.

Banks, J., R. Blundell and A. Lewbel (1997). “Quadratic Engel Curves and Consumer Demand” *Review of Economics and Statistics* 79: 527-539.

Chavas, J.P. and M.T. Holt (1996). “Economic Behavior Under Uncertainty: A Joint Analysis of Risk Preferences and Technology” *Review of Economics and Statistics* 78: 329-335.

Deaton. A. and J. Muellbauer (1980). “An Almost Ideal Demand System” *American Economic Review* 70: 312-326.

DiNardo, J. (2008). "Natural Experiments and Quasi-Natural Experiments". In Durlauf, S.N. and Blume, L.E. (Editors). *The New Palgrave Dictionary of Economics* (Second Edition). Palgrave Macmillan.

Duflo, E., R. Glennerster and M. Kremer (2007). “Using Randomization in Development Economics Research: A Toolkit” in *Handbook of Development Economics* (4): 3895-3962.

Griliches, Z. (1957). “Hybrid Corn: An Exploration in the Economics of Technological Change” *Econometrica* 25: 501-522.

Greene, W.H. (2007). *Econometric Analysis* (Sixth Edition), Prentice Hall.

Hansen, L. (1982). “Large Sample Properties of the Generalized Method of Moments Estimators” *Econometrica* 50: 1029-1054.

Hausman, J. and W. Taylor (1981). “Panel Data and Unobservable Individual Effects” *Econometrica* 49: 1377-1398.

Heckman, J. (1979). “Sample Selection Bias as a Specification Error” *Econometrica* 47: 153–161.

Huffman, W.E. (1977). “Allocative Efficiency: The Role of Human Capital” *Quarterly Journal of Economics* 91: 59-79.

McFadden, R. and K. Train (2000). “Mixed MNL Models for Discrete Response” *Journal of Applied Econometrics* 15: 447-470.

Mundlak, Y. (1978). “On the Pooling of Time Series and Cross Sectional Data” *Econometrica* 56: 69-86.

Smith, V.L. (2008). "Experimental Methods in Economics". In Durlauf, S.N. And Blume, L.E. (Editors). *The New Palgrave Dictionary of Economics* (Second Edition). Palgrave Macmillan.

Wooldridge, J.M. (2002). *Econometric Analysis of Cross-Section and Panel Data* MIT Press, Cambridge, MA.