Chapter 16

Using Survey Data: Some Cautionary Tales

Summary

This chapter illustrates some of the major problems of data quality that every analyst has to confront. These problems are presented in the form of the following 11 "cautions":

1. Do the sampling right.
2. Use a consistent recall method.
3. Use a consistent recall period.
4. Remember that price indexes matter (a lot).
5. Use consistent questions.
6. Adjust for nonresponse bias (if possible).
7. Define expenditure consistently.
8. Value own-farm income properly.
9. Distinguish between values that are zero and those that are missing.
10. Use expenditure per capita, not expenditure per household.
11. Use weights when they are needed.

Be alert to extravagant claims of large jumps in poverty or inequality, and retain a healthy skepticism rather than cynicism. Above all, know your data.

Learning Objectives

After completing the chapter on *Using Survey Data: Some Cautionary Tales,* you should be able to

1. Illustrate what can happen if one uses partial or incomplete samples of data, and if there is nonresponse bias.
2. Explain why it is important to have a consistent method and period of data recall, and why the questions used must be consistent from survey to survey.
3. Show how price indexes are used to inflate poverty lines over time, and how such indexes can be constructed using unit value data.
4. Summarize the strengths and weaknesses four main price indexes—Laspeyres, Paasche, Fisher's Ideal, and Törnqvst—and explain how each is computed.
5. Illustrate the problems that can arise if own-farm income is incorrectly valued.
6. Explain why one cannot, as a general rule, substitute zeros for missing values.
7. Justify the use of expenditure per capita, rather than expenditure per household, in the measurement of poverty and inequality.
8. Explain why sampling weights must be used in almost all cases.
9. React with healthy skepticism to claims of large changes in poverty rates, and analyze the possible causes of such changes.

Introduction: Interpreting Survey Data

This is a chapter of cautionary tales. It provides a series of examples that illustrate how slippery the interpretation of survey data can be, and it draws heavily on examples that Shaohua Chen of the World Bank has compiled based on a decade and a half of working closely with datasets from more than a hundred countries.

The main theme of the chapter is that users of data must be alert to extravagant claims, such as large jumps in income or huge drops in inequality, because these claims rarely hold up to scrutiny. The antidote is to ask questions; before basing conclusions on survey data, it is essential to know enough about how the sample was chosen, how the questions were posed, and how the results were compiled. The examples in this chapter help us to ask the right questions.

Conversely, although we need to be skeptical, there is no reason to become completely cynical about survey data. Some data are unusable some of the time, but this does not imply that survey data are never informative. The examples in

this chapter show how, with care, one can often draw useful conclusions from data that at first sight appear flawed, and how, with care, the quality of survey data can be improved.

Caution 1. Do the Sampling Right

The information in table 16.1 comes from household surveys undertaken in Malawi in 1997 and 2004. The numbers show remarkable progress—a 21 percent increase in mean income, an enormous reduction in inequality, and a sharp drop in the headcount poverty rate. And this occurred during a period when the Malawi's real gross domestic product (GDP) per capita actually fell by 1.3 percent per year.

Are these numbers too good to be true? The answer is yes! For some reason—perhaps due to missing information—more than 3,000 households were dropped from the 1997/98 sample when computing the statistics for that year; the reported sample is 6,586 households in 1997/98 and 11,280 in 2004–05. If the households were dropped randomly, then the results in table 16.1 may still be usable, but there is a suspicion that relatively high-income households were more likely to have been excluded. This would imply that the 1997/98 numbers understate income and overstate poverty, and that the improvement in incomes and poverty incidence shown in table 16.1 are overstated.

Quite apart from the sampling problem, it turns out that the questions used in the 2004–05 survey were more extensive than those used in the 1997/98 survey, so that the measures of expenditure cannot be directly compared between the two. An effort has been made to impute expenditure per capita for items that were not included in the 1997/98 questionnaire. To see how this works, suppose that only the latter questionnaire asked households how much they spent on shoes; then, using the data from the 2004–05 survey, one could regress the value of spending on shoes against other household characteristics, such as total spending on food, the number of adult members, and so on. This estimated equation can then be applied to the data from the 1997/98 survey to generate the predicted ("imputed") value of spending on shoes. The process generates a poverty rate of 53.9 percent for 1997/98, which is more directly comparable to the 52.4 percent rate actually observed in 2004–05

Table 16.1 Income, Poverty, and Inequality in Malawi, 1997/98 and 2004

Survey year	Mean income per person per year, Malawian kwacha (prices of 1997/98)	Gini coefficient	Headcount poverty index (%)
1997/98	399.2	0.503	65.9
2004/2005	483.2	0.390	52.4

Sources: PovcalNet (at http://iresearch.worldbank.org/PovcalNet/doc/MWI.htm); Government of Malawi (2006).

(Government of Malawi 2006). Based on these adjusted numbers, one can conclude that the poverty rate barely fell between 1997/98 and 2004–05.

Caution 2. Use a Consistent Recall Method

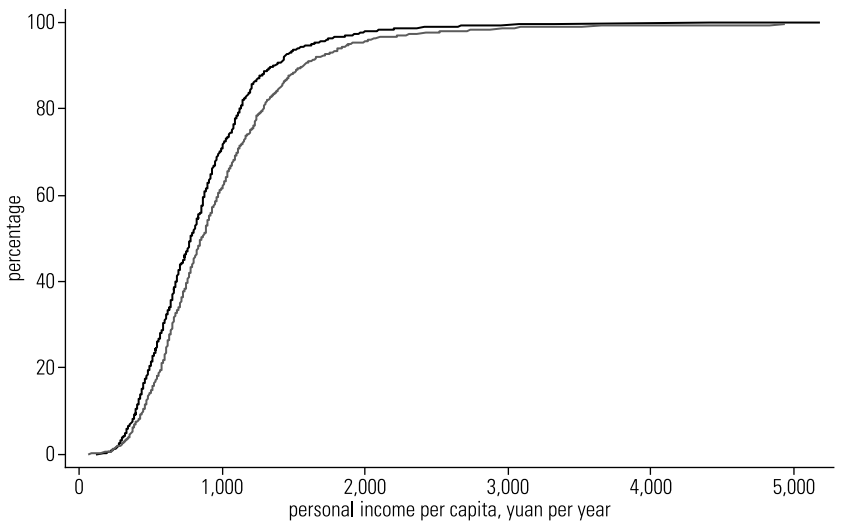
A poverty monitoring survey undertaken in the southwest of China in 1995–96 yielded the following information:

- 1995 mean income per capita: 855 yuan
- 1996 mean income per capita: 993 yuan.

This represents a 16 percent increase in per capita income in one year. Even by Chinese standards, this is a rapid increase—too rapid to be plausible. The cumulative distribution curves for the two surveys are shown in figure 16.1.

Part of the explanation is due to the fact that the two surveys used different methods to collect information on income and expenditures. The 1995 survey used a one-time recall method, whereas the 1996 survey required households to keep daily diaries. It is well know that when the questions are more detailed, or when information is recorded in a more timely fashion, the amounts (spent or earned)

Figure 16.1 Cumulative Distribution Functions, Southwest China Poverty Monitoring Survey



Source: Shaohua Chen, personal communication.
Note: Date for 1995 (upper line) and 1996 (lower line).

will be greater. Of course, this leaves us with a serious problem: there is no easy way to determine how much of the observed 16 percent rise in income was due to a real increase in income and how much to the change in survey procedures.

Caution 3. Use a Consistent Recall Period

Officially, the headcount poverty rate in India fell from 39 percent in 1987–88 to 36 percent in 1993–94 and then dropped dramatically to 26 percent in 1999–2000. The latter results were quickly criticized: was it really possible that poverty fell so rapidly after a period of only modest decline? The surprise was all the greater because an unofficial estimate of poverty, based on a "thin" (that is, smaller scale) survey undertaken in 1998, had shown a poverty rate of 39 percent.

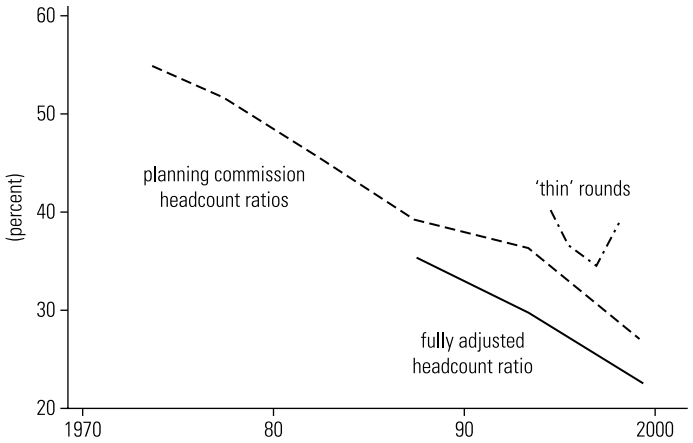
It turns out that there was a technical problem with the results. Indian poverty rates are based on data from the National Sample Surveys (NSS), including the large and important 50th round of 1993–94 and 55th round of 1999–2000. Unfortunately, the expenditure modules of these two surveys are not strictly comparable, for three main reasons:

- The 1999–2000 survey distinguished 173 separate consumption items, which was somewhat fewer than the earlier survey. When there are fewer items, consumption tends to be understated.
- The 1999–2000 survey used a shorter recall period—7 rather than 30 days—for high-frequency consumption items (such as rice and lentils).[1] This would tend to increase measured consumption.
- The 1999–2000 survey used a longer recall period—a year instead of 30 days—for a number of low-frequency items, such as clothing.

The net effect of these changes on reported expenditure levels (relative to a situation in which the same questionnaire were used) is not clear a priori, and thus it is an empirical matter. However, in the case of some medium-frequency items, both surveys used the same questions and the same recall period. Deaton (2001) has found that spending on these items is a good predictor of poverty and has used this to impute the levels of spending on those items where the questionnaire was changed.

Deaton (2001) estimates that if there had been no change in the survey instrument, the official measure of poverty in 1999–2000 would have been 28 percent instead of the reported 26 percent. This represents a considerable degree of poverty reduction, but it is not as striking as the official numbers. And it is clear in retrospect that the results of the 1998 "thin" round of the National Sample Survey were incorrect (mainly because it collected data over a period of just half a year), as figure 16.2 (from Deaton 2002) shows.

Figure 16.2 Headcount Poverty Rates in India, 1970–2000



Source: Deaton 2002.

Note: The "thin" rounds refer to unofficial calculations based on small-scale household surveys, the central line shows the official results, and the lower line shows the calculations made by Deaton.

Caution 4. Remember That Price Indexes Matter (a lot)

There is another important lesson from the Indian experience with measuring poverty: it is essential to use the right price indexes.

Deaton (2001) shows that the official price indexes inflated the Indian poverty line too much between 1993–94 and 1999–2000, and this had the effect of understating the reduction in poverty. Moreover, the official price indexes overstate the price differential between rural and urban areas, greatly overstating urban poverty. Thus, poverty in India is lower than previously had been believed and is falling relatively quickly. This shows up in the "fully adjusted headcount ratio" in figure 16.2.

It is worth looking at Deaton's price calculations in more detail, because they both illustrate good practice and serve as templates for researchers who are trying to measure poverty. To recap, the problem to be tackled is adjusting a poverty line so that it reflects the same cost of living over time. In the Indian case, Deaton took as a given the national poverty line of Rp 115.70 per 30 days per person in 1987–88, but needed to find the equivalent poverty lines for 1993–94 and 1999–2000.

This is the challenge of constructing a price index. The Indian National Sample Surveys (NSS) collect information on the value of expenditures, and the associated quantities of purchases, for many items for each household. This allows one to compute unit values (that is, value divided by quantity), which are akin to prices. Not all of the items are well defined, however. For instance, in the 1999–2000 round of the Indian NSS, "other milk products" included both yogurt (inexpensive) and many

sweets (expensive) and thus was not homogeneous enough to yield a unit value. In practice, Deaton (2001) was able to use items representing three-fifths of household expenditure in the construction of price indexes.

Given unit values (or prices), the next problem is aggregating them into a single price index. There are four common ways to do this (United Nations 2006), and the details are set out in box 16.1. Using data from the Indian NSS of 1993–94 and 1999–2000, Deaton constructed the price indexes that are set out in table 16.2. Panel A shows his estimates of price inflation between 1993–94 and 1999–2000 for rural and urban areas; he estimates that prices rose by about 4 percent less than shown by the official numbers. This implies a lower poverty line than the one used by the official measure, and hence a somewhat lower poverty rate.

Table 16.2, panel B, shows Deaton's estimates of the cost of living in urban, relative to rural, areas. He finds that the urban cost of living is about 15 percent higher, a differential that is less than half of the 37 to 39 percent reported by the official statistics.

The implications of these price adjustments for the measured (headcount) poverty rate in India, both for the 1987–88 to 1993–94 period and for the 1993–94 to 1999–2000 period, are shown in table 16.3. The rural poverty rate in 1987–88 is, essentially by construction, the same according to the official estimate and using Deaton's estimate, but Deaton's measure of urban poverty is far below the rate

Table 16.2 Price Indexes for Inflating Poverty Lines in India

	Budget shares in index						
	1993–94	1999–2000	Laspeyres	Paasche	Fisher Ideal	Törnqvst	Official
Panel A. Prices in 1999–2000 relative to those in 1993–94							
Rural	65.5	59.6	156.4	152.5	154.5	154.5	159.1
Urban	57.8	50.3	162.0	155.7	158.8	157.7	161.4
Panel B. Prices in urban areas relative to those in rural areas							
1993–94	65.8	73.7	117.5	113.7	115.6	115.6	136.7
1999–2000	53.2	61.1	115.9	114.0	115.0	115.1	138.6

Source: Deaton 2001.

Table 16.3 Headcount Poverty Rates for India, Official and Adjusted

		1987–88	1993–94	1999–2000
Official	Rural	39.4	37.1	27.0
	Urban	39.1	33.2	23.5
Price-adjusted	Rural	39.0	32.9	21.6
	Urban	22.8	18.1	9.5
Price- and questionnaire-adjusted	Rural			25.3
	Urban			12.5

Source: Deaton 2001.

Box 16.1 Constructing Price Indexes

There are four popular methods for constructing prices indexes; the details are set out in this box.

1. The **Laspeyres "base-weighted" price index** combines prices by giving each price a weight equivalent to the share of the good in the initial expenditure "basket."

   Formally, let v_{it} be the value of spending on good i in time t, where we typically count $t = 0$ in the base period. And let p_{it} be the price (or, in this context, the unit value) of good i in time t, and q_{it} be the associated quantity. Then the Laspeyres price index is given by

$$L_p = \sum_i \left(\frac{v_{i0}}{\sum_i v_{i0}} \right) \left(\frac{p_{it}}{p_{i0}} \right) = \frac{\sum_i p_{it} q_{i0}}{\sum_i p_{i0} q_{i0}}.$$

   This is probably the most widely used of all price indexes, because it is relatively straightforward to compute. Price data are needed for each period, and spending (or quantity) data are needed just for the base period, which makes it economical when computing regular measures of inflation, such as monthly data on a consumer price index. It generally overstates the "true" inflation in the cost of living, however, because it does not adjust for the fact that consumers substitute away from goods that become relatively expensive, and thus it retains an excessive weight on items that, over time, decline in relative importance.

2. The **Paasche "end-weighted" price index** combines prices by giving each price a weight equivalent to the share of the good in the end-of-period basket. Formally, it is given by

$$P_p = \sum_i \left(\frac{v_{it}}{\sum_i v_{it}} \right) \left(\frac{p_{it}}{p_{i0}} \right) = \frac{\sum_i p_{it} q_{it}}{\sum_i p_{i0} q_{it}}.$$

   The Paasche index overcorrects for substitution effects, and so it generally is held to understate "true" inflation. It generally is not used to construct monthly or quarterly measures of consumer prices because of the expense of regularly updating the detailed information on the value of spending (or quantities bought).

3. **Fisher's ideal index** is the geometric mean of the Laspeyres and Paasche indexes, and so is given by

$$F_p = \left(L_p \times P_p \right)^{1/2}$$

Box 16.1 continued

It can be shown that if households have quadratic utility functions, then Fisher's ideal index generates a "true" cost-of-living index. A practical attraction of F_p is that it is between L_p and P_p, and so avoids the extremes of overstating or understating "true" inflation. To measure Fisher's Ideal Index, information on consumption patterns is needed both in the base year and in the end year, which is impractical (or at least expensive) if one is constructing monthly price indexes, but it is feasible when comparing household surveys at two points in time.

4. The **Törnqvst index** weights the price increases of each good *i* by the budget share of the good averaged between the beginning and end period. This gives

$$T_p = \prod \left\{ \left(\frac{p_{it}}{p_{i0}} \right)^{(1/2)(s_0 + s_t)} \right\} \text{ where } s_{it} \equiv \frac{v_{it}}{\sum_i v_{it}}.$$

If the underlying utility function is logarithmic, then it can be shown that the Törnqvst index measures the true cost of living. Because many economists believe that the utility function is at least approximately (or locally) logarithmic, this is a strength of the index and explains why it is increasingly widely used. In practice, the Törnqvst index usually gives a measure of inflation that is close to the one generated by Fisher's ideal index. Both of these indexes have equally heavy data requirements.

reported officially, and is far more plausible too. Between 1987–88 and 1993–94 the official numbers show a modest reduction in poverty, whereas Deaton finds a more substantial drop (because he uses a price index that inflates the poverty line by less than the official one).

The official figures show a sharp drop in poverty between 1993–94 and 1999–2000. Deaton, using his price deflator, also finds a substantial drop (see the price-adjusted numbers in table 16.3), but he revises these upward somewhat to take into account the effect of changes in the questionnaire used. Deaton's preferred numbers, marked in boldface in table 16.3, show a relatively rapid fall in poverty between 1987–88 and 1999–2000, without any clear change in trend.

Review Questions

1. Sampling bias is common, but can be corrected for relatively easily.
○ True ○ False

2. Which of the following is correct?

- A. The Laspeyres price index typically overstates inflation while the Paasche price index typically understates it.
- B. The Laspeyres price index typically overstates inflation while the Törnqvst price index typically understates it.
- C. The Laspeyres price index typically understates inflation while the Paasche price index typically overstates it.
- D. The Laspeyres price index typically understates inflation while the Törnqvst price index typically understates it.

3. If households are asked to recall their spending levels over a longer period of time, then they will typically estimate annualized spending to be

- Higher
- Lower

4. We have the following information about household purchases of bread and milk in 2007 and 2008:

	Bread		Milk	
	Price	Quantity	Price	Quantity
2007	1.00	4	0.50	3
2008	1.30	3	0.50	4

Based on this information, inflation between 2007 and 2008 was 20.0 percent according to the

- A. Laspeyres index.
- B. Paasche index.
- C. Fisher ideal index.
- D. Törnqvst index.

Caution 5. Use Consistent Questions

When measuring changes in income or spending over time, it is important that the data come from surveys that used comparable questionnaires. The point is obvious enough, but it is worth emphasizing.

The first example comes from Honduras, where a 2003 survey found a headcount poverty rate of 13.8 percent, based on income. Naturally, one wants to compare this with data from previous surveys. The relevant numbers are shown in table 16.4. Are we to conclude that, between 1999 and 2003, poverty fell sharply from 26.3 percent? Or did it rise substantially from 10.7 percent? Or can we not tell? The answer depends on whether the income module used in 2003 was the same as "income module 1" (see table 16.4) or "income module 2." It turns out that the 2003 survey used "income

Table 16.4 Headcount Poverty Rates in Honduras, 1997, 1999, and 2003

	Income module 1	Income module 2	Which income module?
1997	24.1	12.0	
1999	26.3	10.7	
2003			13.8

Source: Shaohua Chen, personal communication.

Table 16.5 Income, Headcount Poverty, and Inequality, Ethiopia, 1999–2000

	Sample size	Mean expenditure per person (birr per month)	Headcount poverty index	Gini coefficient
Welfare Monitoring Survey (June–August 1999)	25,917	46.0	81.3	0.490
Household Income and Expenditure Survey (January–February 2000)	16,982	92.5	21.9	0.300

Sources: PovcalNet, at http://surveynetwork.org/home/?lvl1=activities&lvl2=catalog&lvl3=surveys&ihsn=231-2000-001; and Shaohua Chen, personal communication.

module 2" and thus poverty appears to have risen. It would be easy to make the wrong comparison, however, and erroneously conclude that poverty had fallen sharply.

Our second example comes from two surveys that were undertaken in Ethiopia in 1999–2000 (see table 16.5). According to the Welfare Monitoring Survey, undertaken in June-August 1999, expenditure per capita was relatively unequally distributed (a Gini coefficient of 0.49) and the headcount poverty rate rather high (at 81 percent). But the Household Income and Expenditure Survey (HIES), undertaken in January-February 2000, comes to a different conclusion, finding a remarkably even distribution of per capita expenditure (a Gini coefficient of 0.30) and a headcount poverty rate of just 22 percent.

Why do these results differ so much? Quite simply, the consumption modules used by the two questionnaires differed substantially. It would be necessary to examine each of them in some detail to determine which is more sensible, and which set of results more plausible. The World Bank's PovcalNet reports the data from the HIES, perhaps on the grounds that the expenditure data are more complete and extensive.

Caution 6. Adjust for Nonresponse Bias (if possible)

In very poor countries, compliance rates for surveys are typically high. But as countries become more affluent, it becomes more difficult to persuade people to respond to lengthy questionnaires or to keep diaries. For example, fewer than one in four

people responds to a telephone survey in the United States, which naturally brings into question the representativeness of the results of such surveys.

As long as noncompliance is random, then the survey results are still usable in measuring poverty, income, or expenditure. But it is generally believed that compliance is nonrandom. Richer people are less likely to respond to a questionnaire for a number of reasons:

- Their time is more valuable, so they don't want to spend three hours answering questions.
- They may have more to hide, from the tax collector, for instance, or from prying neighbors.
- They are more likely to have multiple earners in their household, so the information on income is likely to be less reliable, because most surveys only question a single household member about income.
* They are more likely to be away from home.

Some poor people might also not respond to surveys (or be asked to participate), however, if they live in especially remote areas or if they are homeless (and so hard to find, or not on the roster of households), illiterate (especially a challenge when using a diary method to collect information), alienated from society, or illegal residents.

If compliance falls with income, then poverty is overestimated for all measures and poverty lines. It would be useful to be able to correct for noncompliance bias. Consider the basic example set out in table 16.6: a society has two groups of people, the poor (with a 90 percent response rate, constituting 81 percent of those surveyed) and the nonpoor (50 percent response rate, 19 percent of these surveyed). Given these figures, we may infer that 70 percent of the population is poor (= (.81/.9)/(.81/.9 + .19/.5)) and 30 percent nonpoor. Thus, we should weight each observation for the poor by 0.87 (=.70/.81) and each observation for the nonpoor by 1.56 (= .30/.19).

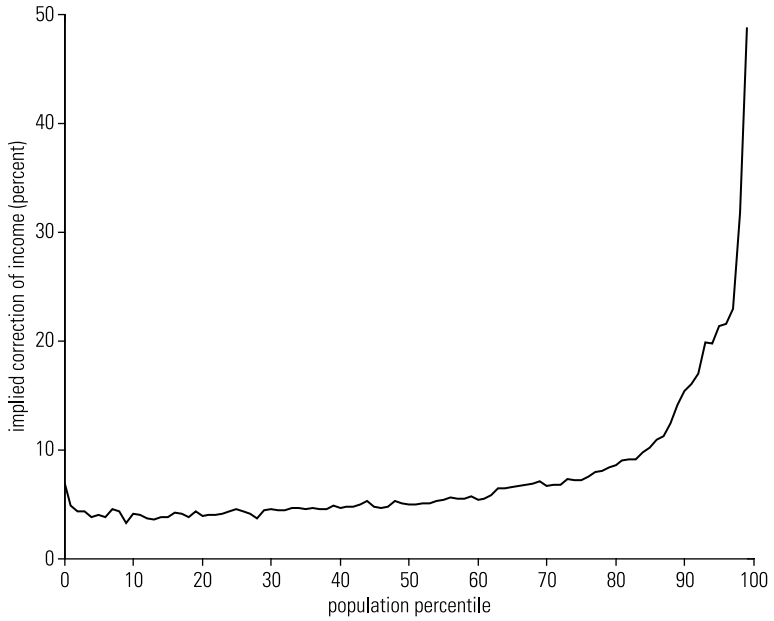
The main practical problem is estimating the response rate, because we do not now whether those who did not respond are rich or poor. An example of the implied correction to income that is needed to adjust for underreporting, for the United States, is shown in figure 16.3.

Table 16.6 Example of Correction for Nonresponse Bias

	"Poor"	"Nonpoor"
Estimated distribution (%)	81	19
But: Response rate (%)	90	50
So: True distribution of population	70	30
Memo: Correction factors	0.87	1.56

Source: Example generated by the authors.

Figure 16.3 Correction Factors for U.S. Income



Source: Shaohua Chen, personal communication.

Caution 7. Define Expenditure Consistently

When making poverty or inequality comparisons, across countries or over time, it is essential that the way in which expenditure (or income) is defined remains unchanged. Yet this is frequently not the case.

Consider the information in table 16.7 for a number of countries in Eastern Europe and the former Soviet Union. The way in which the costs of durable goods and housing are treated varies sharply from one country to the next. It is simply not plausible that durable goods consumption represents 9.5 percent of expenditure in the Russian Federation but only 2.4 percent in neighboring Ukraine or 5.2 percent in Poland. And all of the figures on expenditure on rent appear to be too low—perhaps the imputed costs of housing have not been included. Certainly, it stretches credulity to believe that rent constitutes just 0.3 percent of expenditures in Romania; even the 3.5 percent figure for neighboring Moldova seems modest.

Faced with such faulty numbers, what is the analyst to do? Given enough time and resources, it might be possible to go back to the original data sets to recalculate expenditure for each country on a consistent basis, but rarely do researchers have such a luxury, and besides, this assumes that all of the questionnaires collected adequate information in the first place.

Table 16.7 Percentage of Reported Spending Devoted to Health, Durable Goods, and Rent, for Selected Eastern European and Former Soviet Union Countries, 2002–03

		Percent of consumption expenditure devoted to:		
	Survey year	Health	Durable goods	Rent
Armenia	2002	10.2	0.7	
Azerbaijan	2003	3.2	6.8	0.4
Belarus	2002	2.3	1.6	
Estonia	2003	3.6	6.4	1.1
Georgia	2003	5.3	1.2	0.3
Lithuania	2003	4.1	8.7	
Moldova	2003	5.5	5.1	3.5
Poland	2002	4.7	5.2	3.8
Romania	2003	3.0	1.6	0.3
Russian Federation	2002	2.1	9.5	0.7
Turkey	2002	2.4	7.8	3.3
Ukraine	2003	2.6	2.4	0.4

Source: Shaohua Chen, personal communication.

Table 16.8 Rates of Headcount Poverty and Inequality, with and without Spending on Health, Durable Goods, and Rent, for Selected Eastern European and Former Soviet Union Countries, 2002–03

		Including health, durables, rent		Excluding health, durables, rent	
	Survey year	Poverty rate (%)	Gini coefficient	Poverty rate (%)	Gini coefficient
Armenia	2002	31.1	0.338	33.8	0.292
Azerbaijan	2003	1.7	0.174	5.7	0.181
Belarus	2002	1.4	0.297	1.6	0.292
Estonia	2003	7.5	0.358	9.4	0.339
Georgia	2003	25.3	0.404	27.9	0.403
Lithuania	2003	7.8	0.360	8.7	0.332
Poland	2002	1.9	0.345	2.5	0.324
Romania	2003	12.9	0.311	13.8	0.299
Russian Federation	2002	12.1	0.399	13.6	0.364
Turkey	2002	20.1	0.434	22.7	0.400
Ukraine	2003	4.9	0.281	5.8	0.273

Source: Shaohua Chen, personal communication.

Note: Data use $2 per day standard in 1993 prices.

One approach is to exclude the doubtful expenditure headings entirely. The results of this exercise are shown in table 16.8, where expenditures on health, education, and rent are left out of the measures of consumption. Since no adjustment is made to the poverty line—the World Bank's old $2 per day standard—the reported poverty rates are now higher. In most cases, the Gini coefficient of inequality,

which ranges from 0 for perfect equality to 1 for perfect inequality, falls, implying lower inequality.

In passing, one might note that there are other oddities in the numbers shown in table 16.8: it is strange that the poverty rate in Belarus should be so much lower than in nearby Estonia or Lithuania, which are much richer; and the reported inequality in Azerbaijan is implausibly low.

Is it better to ignore or use bad data? There is no simple answer, but if we are keen to make poverty comparisons, then it would be helpful to develop a set of research protocols that would help ensure consistency in measurement. These protocols would be especially useful in dealing with such problematic expenditure headings as durable goods and the cost of housing.

Caution 8. Value Own-Farm Income Properly

For poor people, a substantial fraction of income (and expenditure) comes from own-farm output. It is therefore important to measure the value of this output correctly if one is to get an accurate measure of poverty.

This is not a trivial point, as the experience with the China Rural Household Survey of 1990 makes clear. The traditional method of imputing income from own-farm consumption used official prices, which by 1990 were far lower than market prices. Using the old method, the headcount poverty rate was 38 percent (see table 16.9), but when own-farm consumption was valued at market prices the headcount rate was just 30 percent—that is, 60 million fewer poor people than had originally been thought!

Caution 9. Distinguish between Values That Are Zero and Those That Are Missing

It is important to distinguish between values that are zero and those that are missing, and it is not generally appropriate to substitute zeros for missing values. For instance, if a questionnaire does not record someone's age, one cannot assume that their age is zero.

Table 16.9 Levels of Income, Inequality, and Poverty in Rural China, 1990

	Mean income (yuan per capita p.a.)	Gini coefficient	Headcount poverty rate (%)
Old method	630	0.315	38
New method	686	0.299	30

Source: China Rural Household Survey 1990.

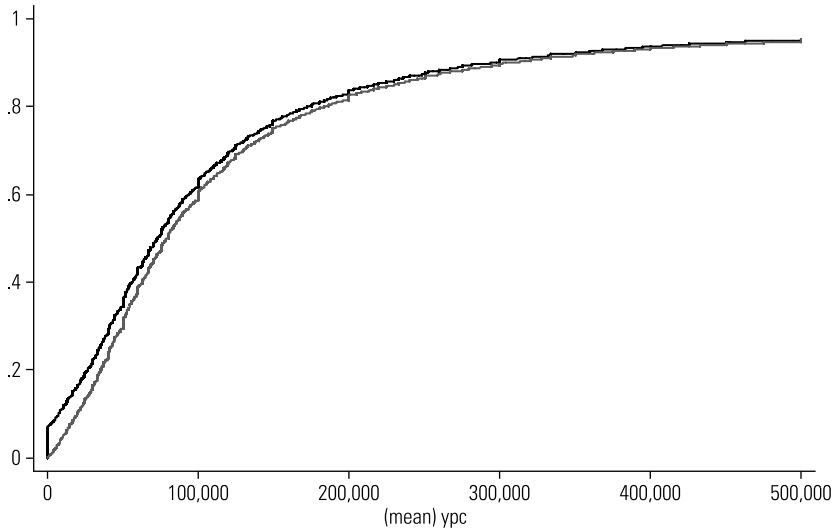
Note: The old method values own-farm production at official prices, while the new method values it at market prices.

The point might seem obvious, but it is sometimes overlooked. In a number of labor force surveys in Latin America and the Caribbean, zeros have been used when information on per capita income was missing. Figure 16.4 shows the cumulative distribution function of per capita income for Colombia for 2003, where 7 percent of the observations on income per capita were missing. When the missing values are included as zeros, one gets the upper curve, which understates the "true" distribution of income. However, the lower line—constructed by excluding the cases with missing values—is a reliable guide to the distribution of income in the population only if missing values are randomly distributed across those surveyed.

There are some occasions in which the use of zeros in the place of missing values may be justified, for instance, if a questionnaire asks the enumerator to fill in nonzero values (for example, for each item of consumption) and to skip an item if it is zero. In such cases, there would need to be a provision for a truly missing value, for instance, using a 99.

Sometimes it is possible to deal with missing values and outliers by going back to the original record to determine whether it is more reasonable to treat a value as zero or missing. This underscores the importance of keeping the original survey records and of putting in place a mechanism whereby the records may be consulted when questions arise about the accuracy of particular numbers.

Figure 16.4 Cumulative Distribution of Income per Capita, Colombia, 2003



Source: Shaohua Chen, personal communication.

Note: The top line shows the distribution when missing values are set to zero; the bottom line excludes observations with missing values.

Caution 10. Use Expenditure per Capita, Not Expenditure per Household

Most surveys collect information on income and expenditure on the basis of the household, rather than on individuals. As discussed in chapter 2, if we are interested in the distribution of welfare, or the poverty rate, we need to rank households by income (or expenditure) *per capita* and not per household.

What happens if the wrong measure is used? Not surprisingly, we may end up drawing the wrong conclusions. This is illustrated in table 16.10, which first ranks individuals by per capita expenditure (correct), and then by per household expenditure (incorrect) based on the Benin Income/Expenditure/Household (IEH) survey of 2003. When ranked correctly, we see that more affluent households tend to be smaller than poor households; but if we mistakenly rank households by income per household, then we are left with the (incorrect) impression that richer households are larger.

Caution 11. Use Weights When They Are Needed

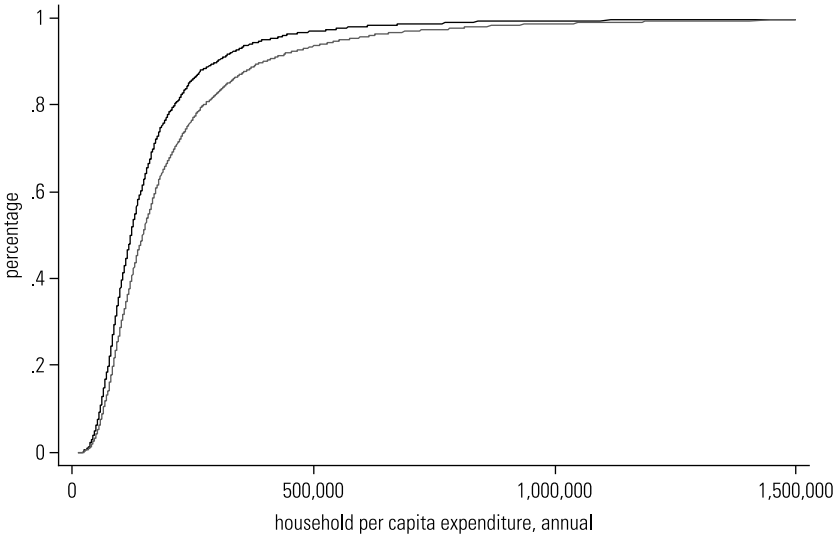
Few living standards surveys are based on simple random samples; most use a stratified cluster sampling design. The implication, as discussed in chapter 2, is that one must use weights when working with the data. The effects of ignoring weights can be striking: the simple average income of tax filers included in the U.S. Internal Revenue Service 1 percent Public Use Sample was $501,814 in 2001—an enormous amount, which reflects the fact that the data set oversamples high-income individuals. When weights are used to adjust for the sample design, one finds an average income, based on the 138,954 observations, of $26,840, which is entirely plausible.

Table 16.10 Household Size by Expenditure per Capita and Expenditure per Household Deciles, Benin, 2003

Households ranked by expenditure per capita		Households ranked by expenditure per household	
Decile	Mean household size	Decile	Mean household size
1 (poor)	8.54	1 (poor)	2.68
2	7.72	2	3.56
3	7.84	3	4.19
4	7.16	4	4.61
5	6.90	5	4.83
6	6.57	6	5.11
7	6.52	7	5.55
8	5.78	8	5.69
9	5.42	9	6.25
10 (rich)	4.33	10 (rich)	6.99

Source: Benin IEH survey of 2003.

Figure 16.5 Cumulative Distribution of Expenditure per Capita, Benin, 2003



Source: Shaohua Chen, personal communication.
Note: Bottom line is incorrect because, unlike the top line, it does not apply sampling weights.

Another example comes from the Benin IEH survey of 2003. Figure 16.5 shows the cumulative density of expenditure per capita both using weights (the upper line, which is correct) and without weights (the lower line). In this particular example, low-income households were undersampled, so when weights were not used, the estimated poverty rate was too low. As shown in figure 16.5, an easy way to see this is to add a poverty line—which would be vertical—and ask which of the distribution curves would show the higher proportion of people in poverty.

Everyone who works with real survey data eventually has a set of stories to tell about the pitfalls that arise easily enough. While this chapter has highlighted some of the more important and obvious problems, there is no substitute for eternal vigilance when working with numbers. This chapter, indeed this entire book, is designed to help get started with data. The more one engages in data analysis using survey data, the more one finds that the topic is both important and fascinating. Enjoy it!

Review Questions

5. Richer people are less likely to respond to questionnaires for all of the following reasons *except*:

- A. Their time is more valuable.
- B. They have fewer members.
- C. They are more likely to be away from home.
- D. They have more to hide.

6. It does not matter what price is used to value own-consumption because it changes both consumption and the poverty line equally.

- ○ True
- ○ False

7. Which of the following statements, derived from the numbers in tables 16.7 and 16.8, is least implausible?

- ○ A. The Gini coefficient is 0.338 in Armenia and 0.174 in neighboring Azerbaijan.
- ○ B. The headcount poverty rate is 7.5 percent in Estonia and 12.1 percent in neighboring Russia.
- ○ C. The percentage of spending devoted to durable goods is 8.7 percent in Lithuania and 1.6 percent in neighboring Belarus.
- ○ D. Housing rental payments constituted 0.3 percent of consumption spending in 2003 in Romania.

8. Send us your own cautionary tale, if necessary with supporting graphs or data. (To Jonathan Haughton, at jhaughto@beaconhill.org.)

Note

1. The 55th round survey asked respondents to give expenditures both for the past 7 days and the past 30 days. But it is widely held that the presence of the question using a seven-day recall is likely to have pulled up reported expenditure levels.

References

Benin, Institut National de la Statistique et de l'Analyse Economique. 2003. "Income/Expenditure/Household Survey of 2003." Cotonou.

China: National Bureau of Statistics Rural Social and Economic Survey Team. Data from *China Rural Household Survey Yearbook 2000*. Beijing: China Statistics Press.

Deaton, Angus. 2001. "Computing Prices and Poverty Rates in India, 1999–2000." Research Program in Development Studies Working Paper, Princeton University, New Jersey.

———. 2002. "Is World Poverty Falling?" *Finance and Development* 39 (2).

Government of Malawi. 2006. *Malawi Growth and Development Strategy: From Poverty To Prosperity 2006–2011*.

PovcalNet. n.d. http://iresearch.worldbank.org/PovcalNet/jsp/CChoiceControl.jsp?WDI_Year=2007.

United Nations Statistics Division. 2006. "Price and Volume Measures." http://unstats.un.org/unsd/sna1993/toclev8.asp?L1=16&L2=3 (accessed December 14, 2006).