SOCIAL STATISTICS: MATHEMATICS MEETS POLICY

by

David S. Moore and Vijay Verma
Purdue University

Technical Report #95-14

Department of Statistics
Purdue University

April 1995

Social Statistics: Mathematics Meets Policy

David S. Moore and Vijay Verma

Revision, 22 February 1995

To appear in S. Garfunkel (ed.), *Mathematics Made Visible*, Springer Verlag.

What percent of the labor force is unemployed? What percent of household budgets is devoted to shelter? How rapidly is the cost of medical care rising? Answers to such questions are staples of news reporting in all developed nations. If the answers are disturbing, the voters will hold the government in power responsible. The technical subject of statistics, which provides the means to produce and interpret trustworthy data, thus affects and is affected by both public policy and electoral politics.

In the Depression era of the 1930s, governments began to intervene more actively in their national economies. To do so, they needed data on employment, prices, and other kinds of economic activity. National statistical systems evolved to provide the data. Statisticians developed an impressive body of techniques to produce data that are accurate, timely, and not impossibly expensive. Statistical sampling methods allow us to estimate the state of an entire population from information gathered from a relatively small sample. We will look at the central ideas of statistical sampling, ideas now widely used in polls of public opinion and in market research as well as in official statistics. It will come as no surprise that applying statistical methods, so straightforward in the pages of a text, to widely dispersed human populations encounters important practical difficulties. We will look briefly at these difficulties as well.

Both the technique and the practice of statistical sampling are highly developed. They lie behind much that is in the news. But the use of technique is a matter of policy, and policy is set by governments whose political fortunes are influenced by the data published by their statistical offices. The Conservative government of Britain changed the official definition of "unemployed" some 30 times during the 1980s. Not surprisingly, almost all of these changes resulted in fewer people being counted as unemployed.[1] Across the channel, the French legislative elections of March 1993, which turned out a Socialist government after a dozen years in office, were heavily influenced by high unemployment. How high? The official figures spoke of 3 million unemployed workers, a rate of 10.5% of the labor force. Just before the elections, the conservative newspaper *Le Figaro* claimed that the "real" unemployment figures were 5.3 million persons and a rate of 21%.[2] *Le Figaro* was not disputing the competence or honesty of the government statisticians, but the policies that decide who is counted as unemployed.

As economic activity becomes ever more international, we need to make reliable international comparisons of economic and social data. There are variations in statistical technique and more important variations in policy among nations that make bald comparison of national data misleading. The unemployment rate in Japan can be compared with that of the United States only with care, and not without qualifications. The Organization for Economic Cooperation and Development (OECD) provides estimates of unemployment in its member states using common methods. In a few cases, OECD figures differ sharply from the official national figures; for example, in mid 1994 the unemployment rate in the Netherlands was 9.9% according to

the OECD and 7.5% according to the Dutch statistical office.

There are even greater difficulties in comparing the income and consumption of households and the standard of living across countries. Within the European Union (EU), however, convergence is taking place in statistics as in other areas. Eurostat, the Union's statistical office, is making special efforts to harmonize the statistics of the member states. Eurostat publishes data based on national statistics which permit international comparisons. The national statistical offices are adopting standards set by Eurostat that allow easier comparison and lessen the risks of political interference. We will therefore use the the EU nations as the leading examples in our discussion.

Policy makers both public and private, business at every level, and informed citizens as well, have a stake in timely and accurate data, easy to interpret and compare across national boundaries. Our topic in this essay is how the science of statistics and the underlying mathematics of probability, in combination with uniform policies and practices, make timely and accurate data possible.

The Art and Science of Sampling

"Must I drink the whole bottle in order to know that the wine is good?" asked the nineteenth century Belgian sociologist and statistician Adolphe Quetelet. The answer is of course no. We often draw conclusions about a whole on the basis of information about a part. That is just what statisticians do when they announce conclusions about all workers or all households after interviews with relatively few among them. A bottle of wine is homogeneous, so that a sip can fairly represent the whole. People and households vary in every important respect, however, so that statisticians face a more difficult task than wine tasters. Variation creates the need for the science of statistics: individuals vary, repeated measurements on the same individual vary, the domain of determinism in nature and everyday life is rather restricted. We must ask how to choose a part from a whole that fairly represents the varying individuals who were not chosen.

The entire group of objects we want to make statements about is called the *population.* A population may consist of people, households, animals, or other objects. It is the target of our desire for knowledge. The individuals from the population from whom we actually gather information form the *sample.*

What shall we measure, and for whom?

The first step in gathering good data about unemployment or household spending, before facing the issue of sampling at all, is to specify exactly what population the data will describe. This simple task presents abundant opportunities to influence the final outcome and to confuse international comparisons. Consider the case of measuring unemployment. The basic population is a nation's labor force, the people able and wanting to work. But what ages are included? If Germany includes only people

aged 15 to 64 and France counts everyone from age 16 up, the two nations' unemployment data refer to different groups of people. What about non-citizens, especially immigrants both legal and illegal from outside the EU? What about students seeking summer employment? What about strikers? There is no single "right" answer to any of these questions, but *some* answer must be given.

Once the target population is firmly in mind, the statistician's next task is to decide exactly what will be measured. Precisely who is "unemployed?" Someone who wants work and doesn't have it, you say. Yes, but ... What if a person wants only part-time work? What if a student wants only temporary work for three summer months? Are people unemployed if they are unwilling to take on the work available because it does not match their needs or qualifications? Again, specific answers must be given, and nations can differ in their choices.

The ambiguity of the term "unemployed" allows governments scope to manipulate unemployment data by social policy. If unemployment is high, offer youth training schemes for the young and retraining programs for middle-aged workers. Allow older workers to draw their government pensions earlier without penalties. Then declare that people in these categories don't count as unemployed. The French call this "traitement social." *Le Figaro* got its higher "real unemployment" figure by adding persons under traitement social to the officially unemployed. Training and other aid programs for the jobless are of course legitimate social policy, but they can have large effects on official unemployment data.

Within a nation, what matters most is simply that the population and the exact quantity measured remain fixed over time. We can then, for example, compare current unemployment rates with those in the past and make clear judgments about the present situation. But comparisons among nations with differing practices can be difficult. As recently as 1987, for example, German and French measures of employment and unemployment differed in these ways:

- Germany considered only persons aged 15 to 64, France everyone over age 16.
- To count as unemployed, a person had to want to work at least 30 hours a week in France. Persons seeking part-time work for at least 19 hours a week could count as unemployed in Germany.
- Persons seeking an apprenticeship in industry were unemployed if they failed to find one in France, but not in Germany.

Other EU nations had still more diverse practices. The author of a survey prepared for the German labor bureau said that, "One has the impression that the differences among nations have grown ever more substantial."[3]

This will never do. A Europe in which citizens of any nation are allowed to work in any other needs consistent data on employment. Eurostat, following guidelines prepared by the International Labor Organization (ILO), produces consistent labor market data for all members of the EU. Nations often keep their existing systems

for comparison with the past, but they also provide Eurostat with the information needed for its work. For example, Eurostat defines the population of working age as all persons aged 14 and above. This population is then divided into three groups—persons in employment, unemployed persons, and inactive persons—on the basis of activity within a specific "reference week." These groups, and some subgroups within them, are defined by specific ILO standards, with minor modifications agreed to by the countries of the Union. Eurostat also prepares a detailed "Union List of Questions" specifying the items of information and exact classifications that each nation is to submit. All this allows Eurostat to produce a standard set of tabulations that make the state of the labor market more comparable across countries.[4]

We will see that not all economic and social data are yet as easily compared as are labor force data. Eurostat's international standards are a model answer to the questions "What is measured?" and "For what population?" Once these questions are answered, we can turn to statistical sampling as the modern technique for producing data about large populations.

Why sample?

Most nations take a *census* of their entire population, often once each decade. The great advantage of a census is that it provides detailed data about every small geographical area in a nation. The United States, for example, has a legal requirement that congressional districts be substantially equal in population. Districts are redrawn after each decennial census, which provides a block-by-block count of the population.

That's one way to measure employment or the breakdown of household expenditures: ask everyone. However, a census is too expensive and time-consuming to be a practical method for producing monthly employment data or annual data about households. Moreover, for most purposes a census is no more accurate than a well-done sample. The best data are produced by interviewers who are carefully trained and supervised. A good interviewer wins the respondent's confidence and corrects misunderstandings to obtain honest and accurate data. Governments can provide trained interviewers for important samples, and now often equip them with laptop computers as well. The computer guides the interview, adjusting the flow of questions to fit earlier responses. If the household size is one, the computer smoothly skips questions about the resident children. A census, on the other hand, must make do with a mixture of mailed questionnaires and personal followup for those who don't respond. The quality of individual responses is likely to be higher in a careful sample survey than in a census.

Administrative records are the source of much national data. People appear in various administrative records as voters, or as taxpayers, or as automobile drivers. For these special populations, the records provide data at very low cost. Denmark, where official registers are comprehensive, has even discontinued its census. In most nations, however, administrative records do not adequately cover the entire population.

Many nations have based their unemployment data on a count of people registered at government employment offices or a count of those claiming unemployment benefits. (The United Kingdom reduced its jobless roll by 180,000 in 1982 by switching from the first to the second of these counts.) Measures of unemployment based on official registers count only people who have some motivation for registering. In the case of taxpayers or automobile drivers, registration is legally compelled. The unemployed are not required to report themselves, however, so data from administrative records are strongly influenced by social policies such as the scope of unemployment benefits. They are susceptible to change when policy changes and are insufficient for international comparisons. And of course administrative records do not exist for most economic and social data because there is no incentive to report.

Sampling, examination of a part in order to obtain information about the whole, is universally employed by governments seeking economic and social data, market researchers exploring consumer tastes, and opinion polls of every variety. We will later describe two important types of governmental sample surveys: labor force surveys and family budget surveys. All such surveys use samples that are small relative to the size of the population. Labor force surveys contact the largest samples, 60,000 to 100,000 households in larger nations. Household budget surveys, which are more intrusive, typically use smaller samples. For example, the Family Expenditure Survey (FES) in the UK contacts 7,000 households each year.

These examples raise an obvious question: How can such relatively small samples provide trustworthy information about large populations. There are about 23 million households in the UK, yet the FES contacts only a few thousand households. The reliability of sample data rests first of all on the method used to select the sample.

How to sample badly

In sampling, as in other areas, the broad and easy way leads to destruction. The easiest sampling design is a *convenience sample* of the part of a population that is most accessible. A convenience sample is analogous to judging the quality of a crate of apples by inspecting a few apples from the top of the crate. These apples may be unrepresentative: apples on the bottom are more likely to be damaged in shipment, and an unscrupulous seller might put the best apples on top. Samples produced by interviewing people in the street are convenience samples. Passers-by in most places and times don't represent the entire adult population. Moreover, interviewers tend to approach well-dressed and non-threatening people, so that the poor and working-class males are systematically underrepresented and the mobile and prosperous middle class is overrepresented in the sample.

A sampling method that systematically overrepresents some parts of the population and underrepresents others is called *biased*. Convenience samples are usually biased. *Self-selected samples*, in which individuals choose themselves for the sample by responding to a general appeal, are also prone to bias. The people who respond

are most often those with some strong motive for doing so, and may be a quite special segment of the population. People with strong opinions, especially strong negative opinions, on an issue are more likely to respond than the general public. Should shops be allowed to open on Sunday? In 1986, when Sunday trading was a live issue in Britain, careful surveys of public opinion showed that between 60% and 70% of British adults favored laws allowing Sunday opening. Yet fewer than 1% of the almost 40,000 letters received by the Home Office were in favor.[5] Letters to government are a form of self-selected sample. They often record strong but atypical feelings.

In both convenience samples and self-selected samples, personal choice produces bias. The statistician's remedy is to allow impersonal chance to choose the sample, so that there is neither favoritism by the sampler nor self-selection by respondents. Choosing a sample by chance eliminates these sources of bias by giving all an equal chance to be chosen. Rich and poor, young and old, brown-skinned and white-skinned, all have the same chance to be in the sample.

Random sampling

The deliberate use of chance in choosing a sample is the central idea of statistical sampling. The simplest way to use chance amounts to placing names in a hat (the population) and drawing out a handful (the sample). This is *simple random sampling.* The formal definition of a simple random sample (SRS) of a given size, call it n, is that it consists of n units from the population chosen in such a way that every set of n units has an equal chance to be the sample actually selected. The point is that simple random sampling not only gives each individual the same chance to be chosen but also gives an equal chance to every possible sample that might be formed. Drawing names from a hat conveys the idea, but in practice computers are good at choosing an SRS from a long list of individuals.

An SRS drawn from a large population is like a list of lottery winners. Everyone has the same chance of winning, but the winners are different every time. The data produced from the sample are therefore also different every time. If we took many simultaneous labor force samples, we would get many different unemployment rates. That's disturbing. To understand why at least some sampling schemes produce trustworthy data, we must look more deeply into the behavior of random sampling. The key to understanding is to ask, "What would happen if we took many samples under the same circumstances?" This is the question that is answered by the mathematics of probability, the same mathematics that applies to lotteries. As in so many instances, mathematics lies just below the surface of important public issues.

Population Information From a Sample

The mathematics of probability describes the patterns that appear when random behavior is repeated many times. It is a remarkable fact about the world that

some phenomena—tossing coins, choosing lottery winners, drawing random samples—display regular patterns when repeated many times, even though individual outcomes are unpredictable. Let us first examine the behavior that probability describes. We will look at a hypothetical sample in France. Our sample will be an SRS for simplicity, and too small for practical use. The principles we will discover, however, apply to all random samples.

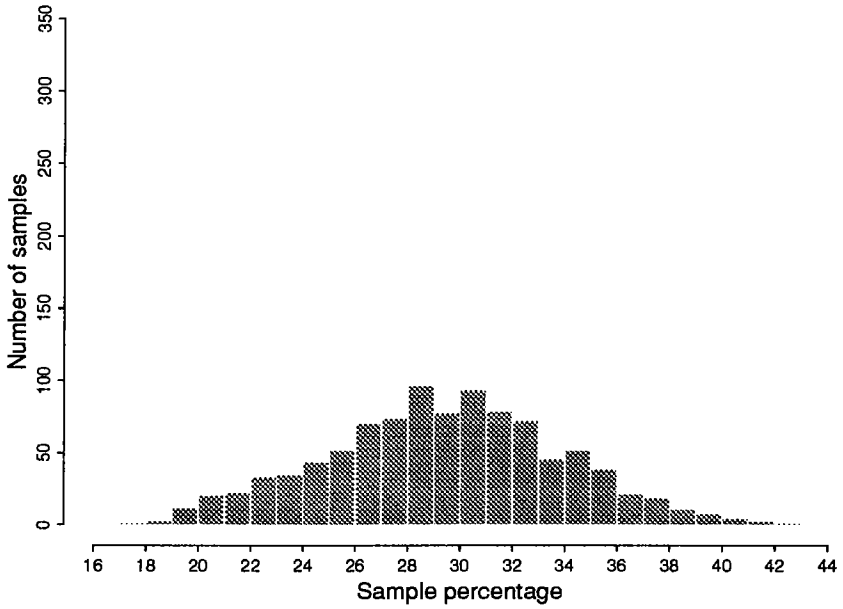
A French sample. The percentage of households consisting of just one person is rising in all developed countries as marriage occurs later and child-bearing declines. We would like to know what the percentage of single-person households is in France. We can't afford to visit every French household, so we will take an SRS. We illustrated the idea of random sampling by the image of drawing names from a hat. Suppose now that we have a very large hat, with an card of the same size and shape for every household in France. Some of the cards are light in color and some are dark. The dark cards represent the single-person households. What percentage of the cards in the hat are dark? We don't know. Looking at all the cards (a census) is beyond our means, so we draw an SRS of size 100.

Our sample contains 35 dark cards. That's 35% of the 100 cards in the sample. Because an SRS gives all cards the same chance to be chosen, it should fairly represent the entire population. So we estimate that 35% of all the cards in the hat are dark. That is a simple version of what French government statisticians do: select a random sample of households and use the sample result to estimate the unknown truth about the population of all households.

The estimation procedure in our example seems fair in the sense that all cards or households have the same chance to be in the sample. But because the sample was chosen by chance, if we repeat the process we will get a different sample and very probably a different percentage of dark cards. Draw another SRS from the hat. This time there are 28 dark cards, so we estimate that 28% of all the cards in the hat are dark. That is, we estimate 35% or 28% depending on the luck of the draw, and drawing more samples will give additional different outcomes. Random sampling eliminates *bias* in choosing a sample, but it certainly does not eliminate *variability.* It is time to ask the key question, "What would happen if we took many samples from this same population?"

Figure 1 displays the percentages of dark cards in 1000 samples, each of size 100, drawn from our French hat. (In fact, we used a computer to simulate the hat and the sampling.) The height of each bar shows the number of samples having a particular outcome. The pattern formed by these outcomes is indeed quite regular. The graph peaks in the center and falls off symmetrically on both sides. But the regular pattern

Figure 1: Sample proportions from 1000 SRSs of size 100



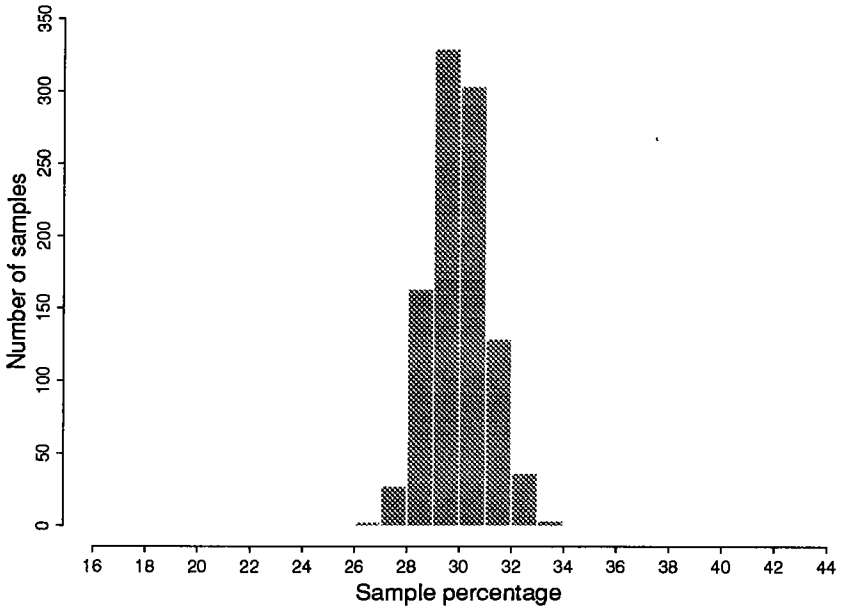
offers little consolation in the presence of so much variation from sample to sample. It is clear from the graph that from one sample of size 100 we might well estimate 22% dark cards or 36% dark cards. That range covers almost the entire range of percentages of single-person households in the EU countries. An SRS of size 100 will not give useful information about France.

A countermeasure lies close at hand. *Larger samples have less variability than small samples.* Let us try our experiment once more, this time drawing SRSs of size 1600 rather than size 100. Figure 2 displays the results of 1000 such samples, on the same scale as Figure 1. The pattern of outcomes is again quite regular and symmetric. But there is now much less variation among the results. It appears that we would almost never estimate fewer than 27% dark cards or more than 33% dark cards if we had just one sample from this hat.

In fact, our hat contains 30% dark cards. (This imitates France, because about 30% of French households consist of just one person.) Look again at Figures 1 and 2. The regular patterns of sample results are centered at 30%. Some samples give estimates that are too high and some give low estimates, but there is no systematic over- or underestimation. The fairness, or lack of bias, of an SRS is reflected in the fact that the distribution of sample results is symmetric and centered at the truth about the population. If the population had 35% single-person households (as Germany does), both graphs would be centered at 35%.

What about variation in the results of repeated samples? The highest and lowest of 1000 estimates are not very informative. In practice, estimates that far off represent very bad luck indeed. Let us rather look at the range covered by the center 95% of

Figure 2: Sample proportions from 1000 SRSs of size 1600

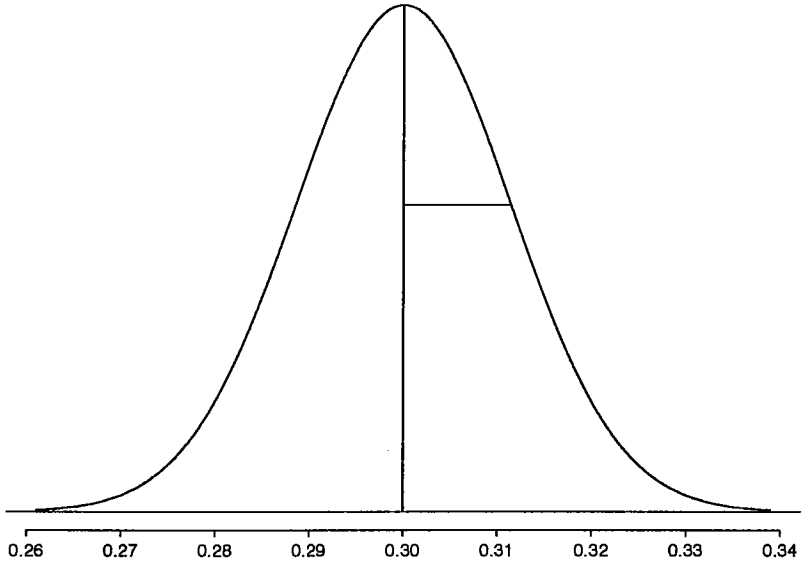


all samples. If we use random sampling repeatedly (as government statistical offices and opinion polls do), we can expect that in the long run 19 out of 20 samples will produce results in this range.

The detailed data behind Figure 1 show that 95% of these 1000 sample outcomes fall between 21% and 39%. We know that 30% is the truth about the population, so write this as $30\% \pm 9\%$. That is, 95% of all SRSs of size 100 give results within 9 percentage points of the truth. That's much too large a spread, so we try samples of size 1600. Now the data show that 95% of our 1000 sample outcomes lie between 27.8% and 32.2%, or $30\% \pm 2.2\%$. Samples of size 1600 produce a result within about 2.2 percentage points of the truth 95% of the time. If this variability in results is still too large, take a yet larger sample. Recall that labor force surveys select tens of thousands of households.

Every answer suggests more questions: Just how rapidly does the variability of sample results go down as we take larger samples? Our hat held 30% dark cards, imitating a population with 30% single-person households. Would the spread of the sample results change if the hat held 20% or 50% dark cards? Does it matter if we are sampling from a small population (Swiss households, say) or a large one (American households)? Answering every question about the behavior of samples by computer simulation would quickly grow tedious. It is time to turn to the mathematics of sampling.

Figure 3: The normal curve for outcomes of SRSs of size 1600



Mathematics to the Rescue

The mathematics of probability describes exactly what the pattern of outcomes of repeated samples must be, if only we repeat the sampling process often enough. In fact, probability describes what would happen if we kept sampling forever, so that the number of samples became infinite. That is, the mathematics is an idealization that describes the real world of sampling only approximately. But the approximation is good enough for practical use (and more advanced mathematics can even tell us how accurate the approximation is). In exchange for idealization, we get generality: we can describe the pattern of outcomes for SRSs of any size from any population.

First, some notation. We are drawing a sample of n items from a population that contains N items. In our French example, N is the number of households in France, about XX million. We drew samples of two sizes, $n = 100$ and $n = 1600$. We want to estimate the proportion of French households that consist of a single person. Call this proportion p. For the simulations in Figures 1 and 2 we took $p = 0.3$, but in practice p is unknown—that's why we need a sample. To estimate the unknown population proportion p, we use the proportion $\hat{p}$ of single-person households in the sample. The sample proportion $\hat{p}$ varies from sample to sample. Our first two samples of size 100 had $\hat{p} = .35$ and $\hat{p} = .28$.

Now we can refine our basic question, "What happens in repeated sampling?" We ask "How does the sample proportion $\hat{p}$ vary when we take repeated SRSs of size n from a population of size N that contains proportion p of the characteristic we are counting?" This isn't an easy question, but mathematicians have had several

centuries to work on it. Here are the facts.

Fact 1. The frequencies with which the possible values of $\hat{p}$ occur in many samples have a very regular pattern. In fact, they follow a *normal curve* like that in Figure 3.

Fact 2. The center of this curve (marked by the vertical line in Figure 3) lies exactly at the true population proportion p.

Fact 3. The amount of variability in the values can be described by the *standard deviation* of the curve. This is the length of the horizontal segment inside the curve in Figure 3. The recipe for the standard deviation of this normal curve is

$$\sigma = \sqrt{\left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n-1}}$$

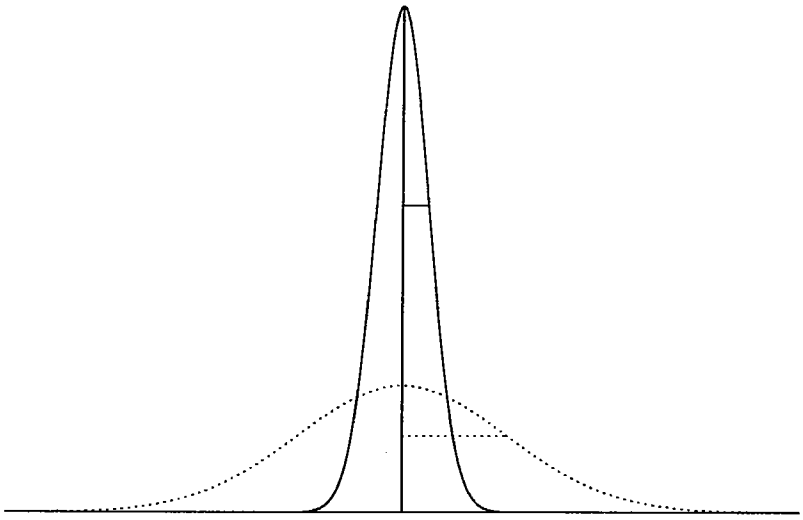
Each of these mathematical facts requires some commentary. Fact 1 replaces observed patterns like those in Figures 1 and 2 by normal curves. Think of the normal curves as arising from drawing a smooth curve through the tops of the bars in Figure 1 and Figure 2. The specific normal curve in Figure 3 matches the experimental results in Figure 2 for repeated SRSs of size $n = 1600$, as you can see by comparing the horizontal scales and overall shapes in the two figures. But the normal curve is not merely an *ad hoc* approximation; it can be *proved* that the values of $\hat{p}$ in many SRSs *must* be very closely described by a normal curve. The only requirement is that the size of the sample must be reasonably large. The normal curve approximation is adequate for our samples of size $n = 100$ and very accurate for sample sizes as large as our $n = 1600$. Of course, the normal curve is an idealization that assumes we sample forever, whereas Figure 2 summarizes a mere 1000 samples.

Figures 1 and 2 have vertical scales that record the count of samples. Because the bars in each figure have equal widths, areas in the figure also represent counts of samples. Figure 3 describes the outcomes of infinitely many samples, so we cannot count outcomes. How shall we choose its vertical scale? Normal curves are scaled so that *areas under the curve are proportions of all possible samples.* The total area under a normal curve is exactly 1. Our eyes, which respond to areas, continue to tell us which outcomes are common and which are rare.

Now we can interpret Fact 2. The normal curve is symmetric, so exactly area 1/2 lies on either side of the center. That is, half of all the values that $\hat{p}$ takes in repeated sampling lie on either side of the center of the curve. Fact 2 tells us that the center lies right at the population proportion p, *whatever this proportion may be.* Although the sample estimate is rarely exactly correct, it has no systematic tendency to either overestimate or underestimate the true population value p. This is what we mean by absence of bias.

Fact 2 says that the sample proportion is "correct on the average" in many samples. Fact 3 allows us to understand and control the variability of $\hat{p}$ about the true p.

Figure 4: Two normal curves with different standard deviations



The distribution of values taken by $\hat{p}$ in many samples always follows a normal curve centered at p. There are many normal curves. All have the famous "bell shape," but the spread of the bell is determined by the standard deviation of the curve. Figure 4 compares two normal curves, one (solid) with a standard deviation one-fourth that of the other (dotted). The curve with the smaller standard deviation is taller and more peaked, representing values that vary less about their center. In both cases, the standard deviation is the length of the horizontal line segment inside the curve. This line segment joins the center line to the point on the curve where the curvature changes from falling-ever-more-steeply as we move away from the center to falling-ever-less-steeply. You can roughly see the standard deviation of any normal curve in this manner.

If the standard deviation is small, almost all values of the sample proportion $\hat{p}$ will lie close to p, and we can be confident that almost all samples will estimate p accurately. The formula for the standard deviation in Fact 3 therefore tells us when we can trust the results of an SRS. Let us examine one by one the factors that appear in that formula,

$$\sigma = \sqrt{\left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n-1}}$$

- $\frac{n}{N}$ is the *sampling fraction,* the proportion of the population (size N) that is in the sample (size n). In the economic and social samples we are interested in, the sampling fraction is very small and the term $1 - \frac{n}{N}$ is very close to 1. This term shows that *the size of the population does not affect the behavior of a sample* as long as the

population is much larger than the sample. A sample of size 1600 will do as well, other things being equal, in the U.S. as in Switzerland. This is good news if you must sample U.S. households, but the Swiss gain nothing from their much smaller population size.

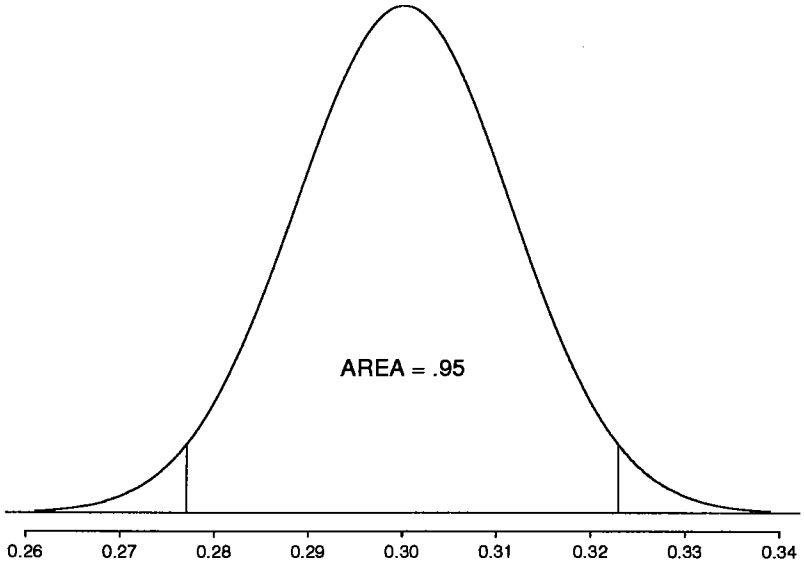
- The population proportion p influences the standard deviation of $\hat{p}$ through the expression $p(1-p)$. This expression is near 0 when p is near either 0 or 1. That makes sense—if almost everyone or almost no one in the population has the characteristic we are recording, there will be little variation from sample to sample. The largest value of $p(1-p)$ is 0.25, attained when $p = .5$, and it changes only slowly as p ranges from about 0.3 to 0.7. The spread of the values of $\hat{p}$ therefore depend a bit on the true p, but this effect is important only when p is close to 0 or 1.
- By far the most important determinant of the standard deviation is the sample size n. If the population is large and we use the conservative value 0.25 for $p(1-p)$, we can guarantee that the standard deviation will be no larger than $0.5/\sqrt{n-1}$. *The variability of the sample decreases as the sample size increases, at a rate proportional to the square root of the sample size.* The two normal curves in Figure 4, one of which has standard deviation one-fourth that of the other, demonstrate the effect of increasing sample size by a factor of 16. This is exactly the comparison between samples of size $n = 100$ and $n = 1600$, first presented in Figures 1 and 2.

Statisticians use the normal curves to study in detail how sample results will vary under different conditions. We can, for example, decide in advance how large a sample we need to estimate facts about the population with the accuracy we desire. Here is one simple fact about normal curves that allows very useful conclusions:

- In any normal curve, 95% of all the observations lie within two standard deviations of the mean.

Figure 5 illustrates this fact for SRSs of size $n = 1600$ drawn from a large population for which $p = .3$. The center of the normal curve that describes the values of the sample proportion $\hat{p}$ in all possible such samples lies at 0.3, the truth about the population. The recipe for the standard deviation (with n/N close to 0) says that σ is about 0.0115. You can verify that this result doesn't change much as p changes; in fact, $\sigma = .0125$ when $p = .5$, and this is its largest value. So when $p = .3$, 95% of all samples of size 1600 give a result that lies between the vertical line segments in Figure 5, which are positioned ± 2.2 from the center. No matter what the unknown values of p may be, 95% of all samples give a $\hat{p}$ no more than 2.5 percentage points away from the population truth. Increasing the size of the sample to 15,000 would produce a sample result guaranteed to be within eight-tenths of one percentage point

Figure 5: The central 95% of observations described by a normal curve lie within two standard deviations of the mean.



of the truth in 95% of all samples. That's close enough to satisfy almost anyone who wants information about French households.

Statistical confidence

The language that is used to announce the results of sampling, whether by government sample surveys or opinion polls, rests on thinking about what would happen in many similar samples. If French government statisticians actually contacted an SRS of 1600 French households and found that this one sample contained 29% single-person households, a brief report might say,

> The survey found 29% single-person households. With 95% confidence, the percentage of single-person households in France is between 26.8% and 31.2%.

Comparing this announcement with our discussion uncovers its meaning. This particular sample gave the result 29%. That is no doubt not exactly correct for the entire population. But 95% of all such samples produce results within $\pm 2.2\%$ of the population truth, so unless we were quite unlucky the true population percentage lies between 26.8% and 31.2%. We can't be *certain* of this, because this sample might be one of the 5% that miss by more. To say that we are *95% confident* that the truth lies between 26.8% and 31.2% is to say "We got this result by a method that gives correct results in 95% of all possible samples." Statistical language translates a probability

statement about what would happen in repeated samples into a statement about how confident we are about conclusions based on a single sample.

News reports often abbreviate yet more: "A government survey says that people living alone make up 29% of French households. The survey's margin of error was 2%." Stating a margin of error reminds the public that statistical samples don't produce exact results, and also gives some indication of how accurate the announced result is likely to be. Adding the level of confidence tells us just how uncertain the announced results are.

Because the margin of error decreases as the square root of the sample size, rather than as the sample size itself, small margins of error generally require quite large samples. Labor force surveys, which must estimate the unemployment rate with only a small error, therefore involve large samples. What is more, governments often want accurate results for regions within the nation. The sample size n for that purpose is only the fraction of the overall sample gathered in the region in question. Regionalism is another force in favor of large samples. Large samples in absolute terms, that is, but still tiny relative to the size of the entire population.

Strata, clusters, and stages

The SRS is the basic random sampling method, and the simplest to understand (remember the hat) and to describe mathematically. In practice, however, samples from large geographically dispersed populations, like the government sample surveys we have mentioned, have a more complex structure. We rarely possess a complete list of the population from which to draw an SRS, and there are also less obvious reasons to use more elaborate sample designs. Imagine, for example, drawing an SRS of size 1000 people from the UK Post Address File. A large computer would not balk at a list containing more than 50 million names, though entering them would be a trial. The 1000 people selected would be scattered across Britain, and therefore expensive to interview in person and difficult to contact even by telephone. Moreover, a sample selected completely by chance may not adequately cover all interesting parts of the population unless it is very large. If we wish to compare household spending in rural and urban areas, for example, we want to ensure that there are a sufficient number of rural households in our sample. Because most households are urban, we will need a large SRS to guarantee enough rural households. We would like to stipulate in advance how the total sample is divided between urban and rural areas.

To do this, many sample surveys restrict random selection by dividing the population into groups of similar units, called *strata*, and then selecting a separate SRS in each stratum. This is a *stratified random sample.* The strata are chosen based on facts known before the sample is taken, such as which areas are urban and which are rural. We can specify how many households from rural areas are in the sample, yet choose them at random to avoid bias.

Stratified sampling has other advantages as well. A stratified design can produce

more exact information than an SRS of the same size by taking advantage of the fact that units in the same stratum are similar to one another. To see this, imagine that all units in each stratum are identical. Then just one unit from each stratum would be enough to completely describe the population. More generally, it can be proved (mathematics again) that a stratified sample whose constituent SRSs are proportional in size to the size of the strata *always* produces a standard deviation no greater than that of an SRS of the same overall size. Of course, a different choice for the sample sizes in the strata may be dictated by the desire to make accurate separate estimates in all strata.

Strata are groups within the population to be sampled, chosen during the design of the sample using information already available. The sample chooses a few elements in each stratum. Groups within the population can be used in quite a different way as well: the sample may contain all members of some groups, while ignoring all others. In this case, the groups are called *clusters* because the final sample is clustered within relatively few of the original groups. We might choose to visit households clustered within only a few of Britain's postal sectors, for example, thus driving down the cost of personal interviews by grouping the sample households in geographical clusters. Use of clusters is almost always combined with selecting the sample in several stages rather than in one grand random choosing.

Such *multistage samples* are common among national samples of households or individuals. For example, the United Kingdom Family Expenditure Survey chooses British households as follows:

Stage 1. Select a random sample of 672 postal sectors in Great Britain. The sample is stratified by geographic region, type of area, and other factors.

Stage 2. After eliminating business and other ineligible addresses, select about 17 addresses at random from the Postal Address File for each sector chosen at Stage 1.

Multistage samples avoid the need for a national list of households or individuals. The FES needs the Postal Address File for only the postal sectors chosen at Stage 1. The final sample is clustered in relatively few postal sectors, so that a single interviewer can easily reach each cluster. As the FES example illustrates, the sample at each stage of a multistage sampling design may be an SRS or a stratified sample.

The results of multistage cluster samples are generally *more* variable than those of an SRS of the same total size. The cost advantage of easier data collection must be balanced against the cost of taking a larger sample to obtain the desired margin of error. It is the business of statisticians to design samples that are effective and economical for particular populations and purposes, and to analyze the data in a way that reflects the design of the sample. Although practical sampling designs usually restrict random selection through stratification and multistage sampling, deliberate use of chance selection is at the heart of statistical sampling. All random sampling

designs can be analyzed mathematically, and all share the essential properties of the SRS.

In broad terms, what the mathematics of probability tells us about sampling is:

- The shape of the pattern of outcomes is determined by the design of the sampling method. Random sampling designs give unbiased results in the sense that the pattern of outcomes is centered at the population truth.
- The variability of sample outcomes—the spread of the pattern—is determined by the size of the sample, and can be made as small as desired by taking a large enough sample.

In short, random samples produce results that are correct on the average and whose deviation from the truth can be controlled by taking a sufficiently large sample.

The Problems of Practice

Translating principles into practice is rarely straightforward. The laws of probability guarantee that random samples behave nicely, but that is no guarantee that actual sample surveys behave so nicely. Random selection eliminates bias in the choice of a sample from a list of the population. When the population consists of people or households, however, accurate information from a sample requires much more than a good sampling design.[6]

To begin, we need an accurate and complete list of the population, called a *sampling frame.* Because such a list is rarely available, most samples suffer from some degree of *undercoverage.* Until recent years, many government surveys in the UK used the Electoral Register as a sampling frame. The Register leaves out all residents who are not entitled to vote, including non-citizens and peers of the realm. Worse, the coverage of those who are entitled to vote is incomplete. The Register tends to miss the young, those who move often, and in general those who are lax about returning the enrollment form. Even an SRS drawn from the Register is therefore biased, because the Register itself systematically underrepresents some parts of the population. The Postal Address Files, which have replaced the Electoral Register in providing sampling frames for government surveys, reduce but do not eliminate undercoverage.

A more serious source of bias in most sample surveys is *nonresponse,* which occurs when a selected individual either cannot be contacted or refuses to cooperate. Nonresponse to nongovernmental surveys often reaches 30% or more, even with careful planning and several callbacks. Nonresponse can also pose difficulties for government surveys. Rates of nonresponse to family budget surveys in the EU range from 8% in Spain through 26% in France and 28% in the UK to 75% in the Netherlands and 89% in Belgium. In Italy, a government survey conducted by telephone found that the percentage of calls that were not answered was 21.4% between January 1 and April 13 and 41.5% during the holiday period July 1 to August 31.[7] Because nonresponse is

much higher in urban areas, many sample surveys substitute other accessible households in the same area to avoid favoring rural areas in the final sample. If accessible people differ from those who are rarely at home or who refuse to answer questions, some bias remains.

In addition, the behavior of the respondent or of the interviewer can cause *response bias* in sample results, especially to questions about sensitive issues. Respondents may hesitate to reveal information about illegal or socially unacceptable behavior. The sample then underestimates the presence of such behavior in the population. Sample surveys in several European countries have concluded that only one or two percent of the male population are active homosexuals. If some homosexuals are reluctant to admit the fact, the true percentage may be higher. An interviewer whose attitude suggests that some answers are more desirable than others will get these answers more often. The race or sex of the interviewer may influence responses to questions about race relations or attitudes toward feminism. Response bias can be reduced by careful training of interviewers and careful supervision to avoid variation among the interviewers. Training and supervision of interviewers are essential parts of good sample survey practice. Responses to questions that do not address sensitive issues are more accurate. Even here, respondents may be unable to provide accurate information if the survey is poorly designed. Answers to questions that ask respondents to recall past events, for example, may be inaccurate because of faulty memory. In particular, many people "telescope" events in the past, bringing them forward in memory to more recent time periods. "Have you visited a doctor in the last six months?" will often draw a "Yes" from someone who last visited a doctor eight months ago.[8]

The *wording of questions* is the most important influence on the answers given to a sample survey. Confusing or leading questions can introduce strong bias, and even minor changes in wording can substantially change a survey's outcome. The U.S. Current Population Survey long asked respondents "Is there a job from which you are on layoff?" with the intent of learning about temporary layoffs from which the worker expected to be called back. Respondents often interpreted "laid off" as a euphemism for "fired," and so answered yes when they had in fact lost their job permanently. The question now adds the phrase "and expect to be called back" for greater clarity.

In 1975, the British government renegotiated the terms for Britain's membership in the Common Market, and then held a national referendum on the issue. An opinion polling organization tried several versions of the question. Presented with the question,

> Do you accept the government's recommendations that the United Kingdom should stay in the Common Market?

those in favor led those opposed by 18.2%. But if the question asked was,

> Should the United Kingdom come out of the Common Market?

the margin between those who wanted to stay in and those who favored withdrawal was only 4.6%. The power to choose the wording is the power to influence the response.

Dealing with these practical difficulties is part of the science of statistical sampling. Users of data produced by sample surveys need to be aware of the potential weaknesses. Good sampling design is not the whole story. If the results are important to you, find out the wording of the questions, the amount of nonresponse, and the date of the survey as well as enquiring about the design.

Some Household Surveys in the European Union

It will be interesting to look briefly at some of the major households surveys currently undertaken in the European Union. This will give us an impression of how different types of surveys are used to meet differing objectives, how national survey practices differ from one country to another, and what can be done in practice to improve comparability of results across countries.

Three major surveys, concerned with the living conditions of households and persons, that are conducted in all countries of the European Union are the labour force survey (LFS), the family budget survey (FBS), and the recently started household panel on income and living conditions (ECHP).

The three sets of surveys provide rather contrasting pictures of the extent to which different countries use similar survey designs and procedures, and how far the information generated is comparable across the EU countries. At one end we have the very diverse family budget surveys. Special steps have to be taken to harmonise the results from these surveys at the stage of analysis and presentation so as to achieve even a limited degree of comparability. At the other end we have the household panel, a truly multinational survey in which the survey design and procedures are standardised to a high degree. This maximises the potential for obtaining comparable results across countries. The labour force surveys fall somewhere in the middle. The national labour force surveys are based on a common approach and in many respects follow similar procedures, which helps in generating comparable data. But there are also important differences in the national practices, limiting the extent of inter-country comparability.

The Labour Force Survey (LFS)

Compiling comparable statistics on employment and unemployment has been a priority task since the very beginning of the European Community (now the European Union). Although employment and unemployment statistics existed in all member states, the sources and definitions used and the methods of data collection differed too much to permit adequate comparison at the Community level. A Community

LFS was organised as early as 1960 to fill this gap. The survey has been periodically revised with the objective of enhancing comparability of labour force statistics between EU countries, and also with other countries in the region as far as possible. From 1983 it was decided to closely follow the "labour force" approach as defined by the International Conference of Labour Statisticians in 1982.[9]

The LFS therefore has the advantage of being based on a common conceptual framework, which enhances the comparability of the results across countries. Apart from this basic framework, various other technical aspects of implementation are laid down in agreement with national statistical institutes. These include the minimum content of the survey, the list of questions and the common coding of responses, and the principal definitions to be applied in the analysis of the results. The survey timing is also synchronised. In all countries, the annual survey extends across all the months of the year, with each household in the sample interviewed once in the course of the year. From one year to the next, a proportion of the sample households remains the same and a proportion is replaced by new random selections. This is called *sample rotation.* Each person above a certain age in the sample households is questioned on his or her economic activity during the week preceding the interview.

The EU nations also agree to minimum sample sizes for the LFS, sample sizes that take into account the size of the country. Recent sample sizes range from 60,000 and 100,000 households in the five largest countries (France, Germany, Italy, Spain, United Kingdom), and 30,000 to 60,000 for most of the remaining. These are the sizes for common, Union-level use of the results. Countries often supplement the sample sizes as well as the survey content to meet national needs. The United Kingdom LFS for instance covered around 240,000 households in 1993, with 60,000 of these to meet the Union requirements. Finally, the Statistical Office of the European Union (Eurostat) processes and disseminates the information forwarded by national statistical institutes in a standard form.

Comparability Does this effort produce perfect comparability of the survey results, or at least of the survey methods, across the countries? Far from it. While the labour force surveys are standardised in terms of the concepts used and the variables generated, they lack standardisation in many other aspects affecting comparability. For instance:

- The survey questionnaires are designed separately in each country. Though the national questionnaires share many common features, they do not follow any standard "blue print."
- The surveys differ in the organisation and methods of data collection—for example, in the use of telephone versus face-to-face personal interviewing.
- The problem of non-response is more serious in some (mostly northern) countries than other (mostly southern) countries. The procedures used for dealing with

non-response are, again, determined at the national level. This is also true of other technical aspects of the statistical analysis of the results.

These differences are to some extent inherent in differing circumstances when the surveys are organised and implemented independently by individual countries. But there remain some unnecessary differences which could be removed (or at least reduced) to improve international comparability without affecting national requirements. The need for comparability does not require that all the aspects of survey methodology be standardised across the countries. We can draw a distinction between *survey aspects*, which should be strictly standardised, and *sampling aspects*, which should be kept flexible.[10]

The survey aspects include the definition of concepts, variables and the survey population, as well as methods of data collection and data analysis. The sampling aspects include sample size and design, which must suit the conditions and requirements of individual national surveys. Standardising the survey aspects aims to control (make similar) systematic errors which occur in the collection of information. If results from different periods and countries are subject to the same sort of bias, that does not distort the picture when we compare the results. By contrast, in deciding on the sampling aspects, we are concerned with reducing variability, and therefore these aspects should differ so as to take advantage of differences among countries. What is required is not identical sampling procedures, but common standards to be followed.

It is therefore quite appropriate that the sampling methods used for the labour force surveys in individual countries differ, and are determined by the national statistical offices on the basis of the technical and administrative facilities in each country. By contrast, a lack of standardisation in the actual questions asked is perhaps the most important aspect limiting comparability of the labour force surveys across the EU countries. Despite the use of a set of common concepts, definitions and classifications, the questions used differ from one country to another in their content, wording, sequencing, and use of specific probes or cues. These differences can have a significant effect on the comparability of the information obtained.[11] A general revision of the LFS questionnaires began in 1992, and it is possible that it will result in more comparable and improved standards.

The Family Budget Surveys (FBS)

Family budget surveys are among the most comprehensive of the household surveys conducted by all member states of the European Union. Through the years, their scope and content has expanded greatly, and the surveys are increasingly required to furnish comparable information across countries of the Union. The national surveys have a common focus: the study of patterns of consumption and living conditions of private households. Apart from quantitative information on consumption, expenditure and income, family budget surveys cover a wide range of topics relating to the households' levels of living. They cover for example:

- basic characteristics of household members, such as age, sex, relationship, martial status, education and employment;
- housing type and size of accommodation, tenure, amenities etc;
- availability of durable goods to the household, such as a dish washer, deep freeze, microwave oven, video recorder;
- information relating to health care costs, use of services and insurance;
- child-care arrangements;
- transport, e.g., car-ownership, cost of private transport, use of public transport;

Family budget surveys generally cover all private households in the national territory of each country. Persons living in institutions, as well as homeless persons, are generally not included in the survey population.

Some countries (e.g., Denmark, Italy, Spain, and the UK) conduct the family budget survey on a continuous basis, in a way similar to the labour force survey. A continuous survey can provide up-to-date information for the monitoring of changes over time. The survey procedures can be established and improved through accumulation of experience, and the interviewers can be retained for longer periods and trained better. On the other hand, continuous or even very frequent surveys are quite burdensome. This is because a family budget survey involves the collection of complex and very detailed information. Considerable resources are required for the collection and processing of such information. It can also be argued that if changes in the patterns of household consumption are not fast, it is not necessary to survey them frequently. It is for such reasons that other EU countries have chosen to undertake a family budget survey only once in five or more years.

The complexity of the survey also means that its sample size cannot be too large; certainly countries use samples much smaller than those used in their labour force surveys. However, in the absence of any common stipulation on sample size, we find quite a wide range of sample sizes used in the family budget surveys in the EU countries. There is a 20-fold variation, from 2,500 households in Denmark to 50,000 in Germany. The range of variation is further increased when we also take into account the differences in survey frequency. Italy for instance surveys 35,000 households each year, and Belgium around 3,000 once every five years. In the middle range are countries like France, Greece, Ireland, and the UK, with sample sizes of 6,000 to 9,000.

Whether the survey is conducted on a continuous basis or only periodically, any one round of the survey has a fairly typical structure. Generally the sample is spread over a whole year to even out the effect of seasonal variation in the households' consumption and expenditure. As an example, let us consider the survey in France. Many of the other family budget surveys in the EU follow similar models, though there are also many important variations.

France carries out a family budget survey once in every 5 years with a sample size of around 9,000 households. A survey extends over 48 weeks (i.e., the whole year, leaving out four holiday weeks). The total sample is divided into 8 equal subsamples, each of which represents the whole country. These subsamples are active one after the other, each over a period of 6 weeks, during the survey year. Such a structure is introduced to ensure that all seasons are covered equally well in the whole country. In the implementation of a subsample, each household in it comes into the survey for a period of two weeks. As in family budget surveys in the other countries, data collection involves a combination of

- one or more personal interviews, during which information of a more general nature is obtained; and
- diaries maintained by households or individuals to record on a daily basis expenditures and other details which are easily forgotten.

During the two weeks a household is in the sample, its members keep diaries to record all items of expenditure. In addition, the household is involved in two main interviews. One takes place immediately before the two weeks of diary-keeping, to obtain information on general characteristics and expenditures. The other occurs immediately after that period and collects information on household and personal income.

So we see that the structure of a family budget survey can be quite complex and the information sought very detailed. And the survey of France is by no means one of the more intensive ones in the European Union. In fact it is, relatively speaking, one of the "light" surveys! In some countries (the Netherlands, Germany, and especially Belgium), the family budget surveys are much more detailed, requiring a year-long participation of each sample household in the survey.

Diversity Notwithstanding their common purpose and features, the family budget surveys in the EU display a wide diversity of designs and methods, as well as major differences in the concepts used and topics covered. These surveys are clearly less uniform than the labour force surveys in the EU countries. Three main factors have contributed to this diversity. Firstly, there is the weight of history, tradition and preference. Generally the family budget surveys have evolved more or less independently as national undertakings. Secondly, internationally accepted standards and guidelines of the type available for labour force statistics, are lacking in the field of household income and expenditure. Thirdly, this diversity reflects the complexity of such surveys, permitting a great variety of arrangements to choose from.

Without going into detail,[12] the family budget surveys in EU differ in their timing and frequency, the definition of household and household membership, and the length, detail, and period of the diaries they require. Belgium asks each sample household to keep a diary for a whole year, while Greece, Spain, and Portugal require only one

week. Comparability is also damaged by the lack of random sampling and/or high rates of non-response in some countries. This happens especially where the survey is too detailed and burdensome, as in Belgium, Germany, and to some extent in the Netherlands. In Germany, the sample is not even designed to be a random sample, a highly undesirable practice in surveys of such national importance.

There are even national differences in the concepts used and variables measured. For some surveys, the objective is to measure what households spend, in others to measure what they actually consume during a reference period. Particular components of consumption or expenditure may be covered in more or less detail. Income is treated very differently in the various surveys. Obtaining good information on income in its own right is an important objective in some surveys (e.g., Netherlands, United Kingdom), but is not considered so central in others (e.g., Ireland). The procedures for obtaining information on income also differ. Belgium and Germany for instance require continuous recording by the household in a diary over the whole year; by contrast in Italy only approximate information is sought on the basis of a single question.

Harmonisation The family budget surveys share a common focus and objectives, and in fact constitute the only source of more or less comparable information for the study of household consumption and many other indicators of conditions of life in the European Union. It is therefore important to make the results more comparable across countries despite major substantive and statistical differences. Eurostat has launched a progressive, step-by-step programme of work to enhance the quality and comparability of the existing surveys, which it terms the "harmonisation of family budget surveys".[13]

Harmonisation has two main thrusts. The first is to *make the best use of the data as collected in the national surveys.* Eurostat periodically publishes an elaborate set of tabulations providing widely-used comparative data on all member countries. We may note in particular that a great deal of research on poverty and relative deprivation draws on these compilations. Differences which still remain are carefully documented, so that they can be taken into consideration in comparing the results. The second emphasis is to *improve the national surveys.* Eurostat offers recommendations for improvements in the timing, content and methodology of the surveys, and for making them more comparable. However, it remains up to the countries to decide to what extent they can implement these recommendations.

Household Panel on Income and Living Conditions (ECHP)

The best way to obtain comparable information is to proceed from a common design and methodology. This is the approach being followed in the EU household panel. A *panel survey* means that the same set of households and persons is followed over a period of time, at the same time updating the sample so that it continues to represent

the population as it changes over time. During 1993, the 12 countries that then made up the EU conducted pilot studies for the establishment of their national panels. Full-scale surveys started in 1994. The survey will cover a wide range of topics on living conditions such as income, employment, housing, health, and so on. From its very conception, ECHP is a *Union-wide* undertaking, a truly multinational survey that covers some non-EU European countries as well. Its specific characteristics and arrangements derive from its European perspective. It is more than a series of co-ordinated national household panels and is conceived to address issues of European policy, in particular the monitoring of social and economic changes following the establishment of a single market after 1992.

In design, the panel survey involves once-a-year interviewing of a nationally representative sample of households and persons aged 16 years and over. These units are followed up each year thereafter. The objective is to assess changes at the individual level within a dynamic framework, for example, to study how households and individuals move in and out of poverty or unemployment and how various events in their personal, family and social life affect each other. The sample sizes are quite similar to those used in the family budget surveys, though here the sizes are much more uniform across the countries. Special techniques have to be used to ensure that the sample remains representative over time. When households or individuals move, they must be followed up to their new location. As children in the sample households reach the age of 16, they must be included as full-fledged respondents in the survey. And so on.

In addition to shared concepts and standards, a central feature of the ECHP project is the use of a common "blue print" questionnaire which serves as the point of departure for all national surveys. The use of a common instrument ensures not only common concepts and content for the surveys, but also a common approach to their implementation. But let us not forget that the requirement of comparability of the information generated does not necessarily imply the need to use identical questionnaires in all countries. On the contrary, because of differing legal and institutional frameworks, different questions may sometimes be required in different countries to obtain the same sort of information.

We began this chapter by noting the relationship between statistics and policy. Here is a sample of policy-relevant issues which the EU Household Panel on Income and Living Conditions has been used to address.[14] The rich potential of such a survey can be seen more clearly on the basis of longitudinal analysis, linking data over time.

- Can households balance their budgets? What proportion of households are currently unable to pay the monthly rent or mortgage, utility bills or hire-purchase commitments? How does the incidence of such problems vary with household income?

- What is the level and composition of social security benefits for various income

groups? To what extent are benefits received because of poverty, and to what extent because people have been making social security contributions?

- For what reasons do people work less than full-time? Is it because of caring obligations, lack of job opportunities, or participation in education or training? What other factors are involved?
- What are working conditions like, and how satisfied are people with them?
- How burdened are women by household chores, bringing up children, and above all proving care to old and incapacitated family members? Do these responsibilities limit their capacity to undertake paid employment? Do their men folk help?
- What proportion and who among the population are suffering from long-term illness or disability? How many are hospitalised and for how long? How frequently are they seen by a doctor? How many working days are lost due to illness?
- To what extent are individuals socially integrated or suffer from social exclusion? Do they participate in social or cultural organisations? How satisfied are they with various aspects of their lives, from work and income to leisure and social relations?

This list, which could be greatly extended, reflects the complexity and stress of modern society. Government policy addresses all of these issues. Sound statistical methods allow the issues to be studied on the basis of accurate data rather than speculation and special pleading.

Notes

1. See *The Economist*, February 14 1987, p. 50 and September 11 1993, p. 15.
2. Jean Mottin, "Chômage: bilans truqués," *Le Figaro*, March 11 1993, p. 2.
3. Heinz Werner, "Arbeitslosigkeit ist nicht gleich Arbeitslosigkeit," *Materialien Aktuell*, Nr. 3, 1987. Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit, Nürnberg, Germany.
4. For details of the ILO standards, see Hussmans, Mehran, and Verma, *Surveys of the Economically Active Population, Employment, Unemployment and Underemployment*, 1990, International Labor Office, Geneva.
5. *The Economist*, March 1 1986, p. 54.

6. For more detail on the material of this section and complete references, see P. E. Converse and M. W. Traugott, "Assessing the accuracy of polls and surveys," *Science*, 234 (1986), pp. 1094–1098.

7. Giuliana Coccia, "An overview of non-response in Italian telephone surveys," *Proceedings* of the 99th Session of the International Statistical Institute, 1993, Book 3, pp. 271–272.

8. For more detail on the limits of memory in surveys, see N. M. Bradburn, L. J. Rips and S. K. Shevell, "Answering autobiographical questions: the impact of memory and inference on surveys," *Science*, 236 (1987), pp. 157–161.

9. These concepts and related survey methodology have been elaborated in the work cited in Note 4.

10. Leslie Kish, "Multipopulation survey designs: five types with seven shared aspects," *International Statistical Review*, 62 (1994), pp. 167–186.

11. On how such differences in the formulation of survey questions can (may) affect comparability of the results obtained, see Van Bastelaer, *Differences in the Measurement in the Labour Force Surveys in the European Community*, and *Differences in the Designs of the Labour Force Surveys in the European Community and Some Consequences*, 1993, The Netherlands Central Bureau of Statistics.

12. A comprehensive review of these methodologies is given in Verma and Gabilondo, *Family Budget Surveys: Methodology and Recommendations*, 1993, Statistical Office of the European Community.

13. These steps are detailed in several Eurostat reports on Harmonisation of Family Budget Surveys. See in particular Verma, *Standardised Variable Lists for Comparative Tabulation* and *Plan for Comparative Tabulation*, 1991, and Pearce and Verma, *Construction of Standard Variables: Illustrations (Spain, France, Belgium, United Kingdom*, 1991, and *The New Tabulation Plan: Illustrations (Spain, France)*, 1992.

14. *European Community Household Panel: Strategy and Policy*, 1993, working paper by ECHP Development Group, Eurostat.