

Introduction to Descriptive Statistics

Learning Objectives

- **Data Collection**
- **Population**
- **Sample**
- **Statistical inference**
- **Types of Data**
- **Sampling schemes**
- **Sample Size**

General procedure for data collection

1. Define the objectives of the problem and proceed to develop the experiment or survey.
2. Define the variables or parameters of interest.
3. Define the procedures of data-collection and measuring techniques. This includes sampling procedures, sample size, and data-measuring devices (questionnaires, telephone interviews, etc.).

Example 1. We may be interested in estimating the average household income in a certain community. In this case, the parameter of interest is the average income of a typical household in the community. To acquire the data, we may send out a questionnaire or conduct a telephone interview. Once we have the data, we may first want to represent the data in graphical or tabular form to better understand its distributional behavior.

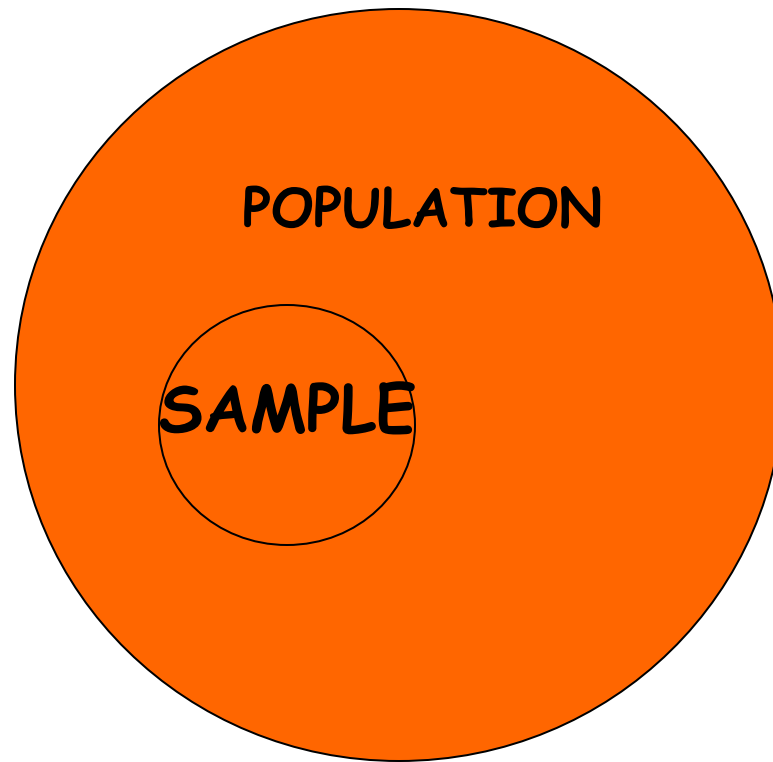
Then we will use appropriate analytical techniques to estimate the parameter(s) of interest, in this case the average household income.

Definition 1. *A **population** is the collection or set of all objects or measurements that are of interest to the collector.*

Example 2. Suppose we wish to study the heights of all female students at a certain university. The population will be the set of the measured heights of all female students in the university. The population is not the set of all female students in the university.

Definition 2. *The **sample** is a subset of data selected from a population. The **size** of a sample is the number of elements in it.*

Example 3. We wish to estimate the percentage of defective parts produced in a factory during a given week (five days) by examining 20 parts produced per day. The parts will be examined each day at randomly chosen times. In this case “all parts produced during the week” is the population and the (100) selected parts for five days constitutes a sample.



Sample and population

Political polls: The population will be all voters, whereas the sample will be the subset of voters we poll.

Laboratory experiment: The population will be all the data we could have collected if we were to repeat the experiment a large number of times (infinite number of times) under the same conditions, whereas the sample will be the data actually collected by the one experiment.

Quality control: The population will be the entire batch of items produced, say, by a machine or by a plant, whereas the sample will be the subset of items we tested.

Clinical studies: The population will be all the patients with the same disease, whereas the sample will be the subset of patients used in the study.

Finance: All common stock listed in stock exchanges such as the New York Stock Exchange, the American Stock Exchanges, and over-the-counter is the population. A collection of 20 randomly picked individual stocks from these exchanges will be a sample.

Definition 3. *A statistical inference is an estimate, a prediction, a decision, or a generalization about the population based on information contained in a sample.*

Example 4. We may be interested in the average indoor radiation level in homes built on reclaimed phosphate mine lands (many of the homes in west-central Florida are built on such lands). In this case, we can collect indoor radiation levels for a random sample of homes selected from this area, and use the data to infer the average indoor radiation level for the entire region. In the Florida Keys, one of the concerns is that the coral reefs are declining because of the prevailing ecosystems. In order to test this, one can randomly select certain reef sites for study and, based on these data, infer whether there is a net increase or decrease in coral reefs in the region. Here the inferential problem could be finding an estimate, such as in the radiation problem, or making a decision, such as in the coral reef problem.

Types of Data

Definition 4. **Quantitative data** *are observations measured on a numerical scale. Non-numerical data that can only be classified into one of the groups of categories are said to be* **qualitative or categorical data**.

Categorical data could be further classified as *nominal data* and *ordinal data*.

Example 5. Data on response to a particular therapy could be classified as no improvement, partial improvement, or complete improvement. These are qualitative data. The number of minority-owned businesses in Florida is quantitative data. The marital status of each person in a statistics class as married or not married is qualitative or categorical data. The number of car accidents in different U.S. cities is quantitative data. The blood group of each person in a community as O, A, B, AB is qualitative data.

Definition 5. Cross-sectional data *are data collected on different elements or variables at the same point in time or for the same period of time.*

Example 6. The data in Table 1. represent U.S. federal support for the mathematical sciences in 1996, in millions of dollars (source: *AMS Notices*). This is an example of cross-sectional data, as the data are collected in onetime period, namely in 1996.

Table 1. Federal Support for the Mathematical Sciences, 1996

Federal agency	Amount
National Science Foundation	91.70
DMS	85.29
Other MPS	4.00
Department of Defense	77.30
AFOSR	16.70
ARO	15.00
DARPA	22.90
NSA	2.50
ONR	20.20
Department of Energy	16.00
University Support	5.50
National Laboratories	10.50
Total, All Agencies	185.00

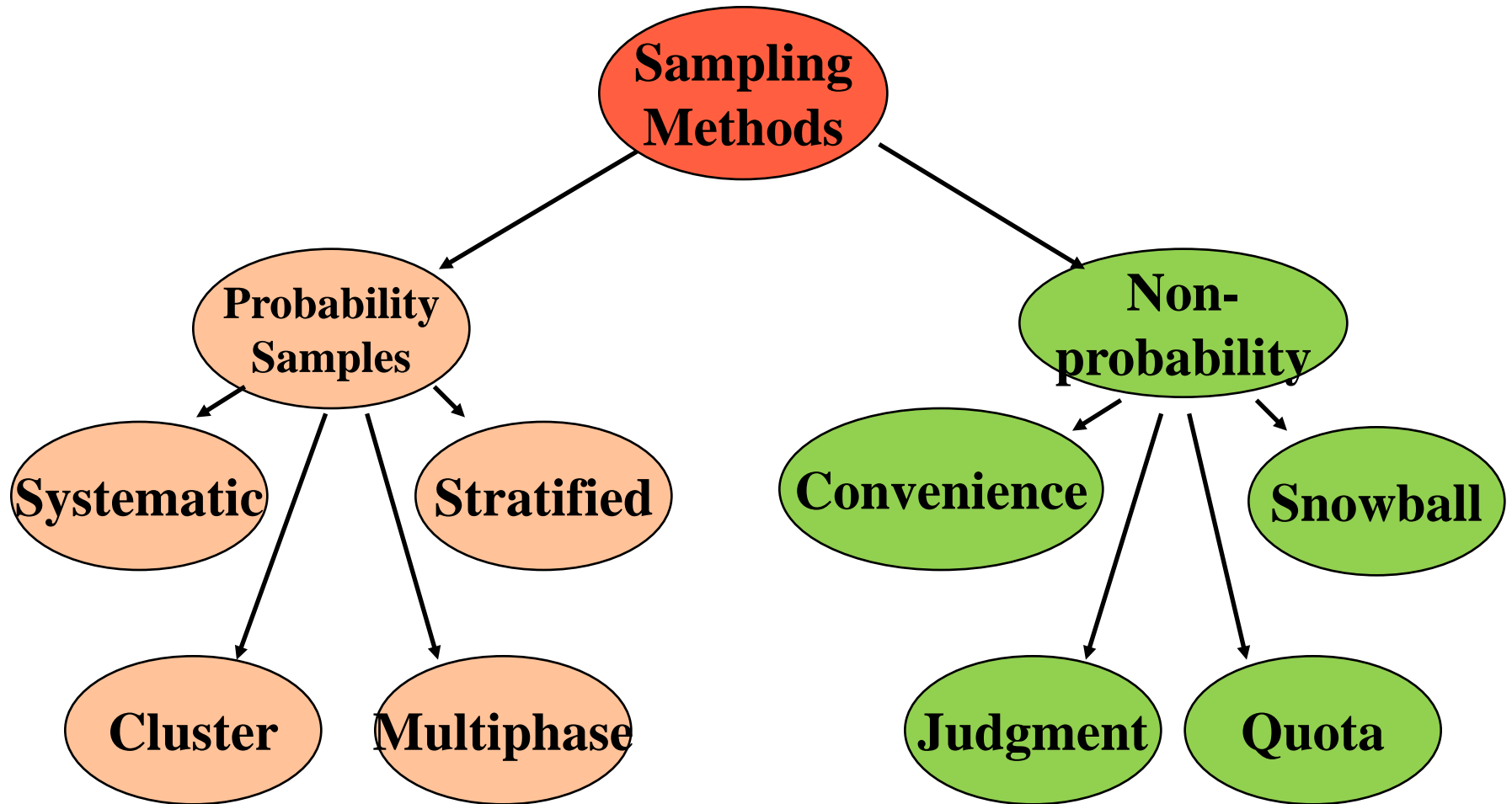
Definition 6. Time series data *are data collected on the same element or the same variable at different points in time or for different periods of time.*

Example 7. The data in Table 2 represent U.S. federal support for the mathematical sciences during the years 1995–1997, in millions of dollars (source: *AMS Notices*) This is an example of time series data, because they have been collected at different time periods, 1995 through 1997.

Table 2. United States Federal Support for the Mathematical Sciences in Different Years

Agency	1995	1996	1997
National Science Foundation	87.69	91.70	98.22
DMS	85.29	87.70	93.22
Other MPS	2.40	4.00	5.00
Department of Defense	77.40	77.30	67.80
AFOSR	17.40	16.70	17.10
ARO	15.00	15.00	13.00
DARPA	21.00	22.90	19.50
NSA	2.50	2.50	2.10
ONR	21.40	20.20	16.10
Department of Energy	15.70	16.00	16.00
University Support	6.20	5.50	5.00
National Laboratories	9.50	10.50	11.00
Total, All Agencies	180.79	185.00	182.02

Classification of Sampling Methods



Probability Samples

Definition 7. *A sample selected in such a way that every element of the population has an equal chance of being chosen is called a **simple random sample**. Equivalently each possible sample of size n has an equal chance of being selected.*

Example 8.

For a state lottery, 52 identical Ping-Pong balls with a number from 1 to 52 painted on each ball are put in a clear plastic bin. A machine thoroughly mixes the balls and then six are selected. The six numbers on the chosen balls are the six lottery numbers that have been selected by a simple random sampling procedure.

Some advantages of simple random sampling

- Selection of sampling observations at random ensures against possible investigator biases.
- Analytic computations are relatively simple, and probabilistic bounds on errors can be computed in many cases.
- It is frequently possible to estimate the sample size for a prescribed error level when designing the sampling procedure.

Definition 8. *A systematic sample is a sample in which every K th element in the sampling frame is selected after a suitable random start for the first element. We list the population elements in some order (say alphabetical) and choose the desired sampling fraction.*

STEPS FOR SELECTING A SYSTEMATIC SAMPLE

1. Number the elements of the population from 1 to N .
2. Decide on the sample size, say n , that we need.
3. Choose $K = N/n$.
4. Randomly select an integer between 1 to K .
5. Then take every K th element.

Example 9. If the population has 1000 elements arranged in some order and we decide to sample 10% (i.e., $N=1000$ and $n=100$), then $K = 1000/100 = 10$. Pick a number at random between 1 and $K = 10$ inclusive, say 3. Then select elements numbered 3, 13, 23, . . . , 993.

Definition 9. *A stratified sample is a modification of simple random sampling and systematic sampling and is designed to obtain a more representative sample, but at the cost of a more complicated procedure. Compared to random sampling, stratified sampling reduces sampling error. A sample obtained by stratifying (dividing into nonoverlapping groups) the sampling frame based on some factor or factors and then selecting some elements from each of the strata is called a stratified sample. Here, a population with N elements is divided into s subpopulations. A sample is drawn from each subpopulation independently. The size of each subpopulation and sample sizes in each subpopulation may vary.*

STEPS FOR SELECTING A STRATIFIED SAMPLE

1. Decide on the relevant stratification factors (sex, age, income, etc.).
2. Divide the entire population into strata (subpopulations) based on the stratification criteria. Sizes of strata may vary.
3. Select the requisite number of units using simple random sampling or systematic sampling from each subpopulation. The requisite number may depend on the subpopulation sizes.

Example 10. In a population of 1000 children from an area school, there are 600 boys and 400 girls. We divide them into strata based on their parents' income as shown in Table 3. This is stratified data.

Example 11. Refer to Example 10. Suppose we decide to sample 100 children from the population of 1000 (that is, 10% of the population). We also choose to sample 10% from each of the categories. For example, we would choose 12 (10% of 120) poor boys; 6 (10% of 60 rich girls) and so forth. This yields Table 4. This particular sampling method is called a proportional stratified sampling.

Table 3. Classification of School Children

	Boys	Girls
Poor	120	240
Middle Class	150	100
Rich	330	60

Table 4. Proportional Stratification of School Children

	Boys	Girls
Poor	12	24
Middle Class	15	10
Rich	33	6

Some uses of stratified sampling

- 1.** In addition to providing information about the whole population, this sampling scheme provides information about the subpopulations, the study of which may be of interest. For example, in a U.S. presidential election, opinion polls by state may be more important in deciding on the electoral college advantage than a national opinion poll.
- 2.** Stratified sampling can be considerably more precise than a simple random sample, because the population is fairly homogeneous within each stratum but there is a sizable variation between the strata.

Definition 10. *In cluster sampling, the sampling unit contains groups of elements called clusters instead of individual elements of the population. A cluster is an intact group naturally available in the field. Unlike the stratified sample where the strata are created by the researcher based on stratification variables, the clusters naturally exist and are not formed by the researcher for data collection. Cluster sampling is also called **area sampling**.*

Example 12. Suppose we wish to select a sample of about 10% from all fifth-grade children of a county. We randomly select 10% of the elementary schools assumed to have approximately the same number of fifth-grade students and select all fifth-grade children from these schools. This is an example of cluster sampling, each cluster being an elementary school that was selected.

Definition 11. **Multiphase sampling** *involves collection of some information from the whole sample and additional information either at the same time or later from subsamples of the whole sample. The multiphase or multistage sampling is basically a combination of the techniques presented earlier.*

Example 13. An investigator in a population census may ask basic questions such as sex, age, or marital status for the whole population, but only 10% of the population may be asked about their level of education or about how many years of mathematics and science education they had.

Errors in Sample Data

In general, there are two types of errors:

- **non-sampling errors**
- **sampling errors.**

It is important for a researcher to be aware of these errors, in particular non-sampling errors, so that they can be either minimised or eliminated from the data collected.

○ **Non-sampling errors**

These are errors that arise during the course of all data collection activities.

In summary, they have the following characteristics:

- exist in both sample surveys and censuses data.
- difficult to measure .

○ **Sampling error**

- ✓ Refer to the difference between the estimate derived from a sample survey and the 'true' value that would result if a census of the whole population were taken under the same conditions.
- ✓ These are errors that arise because data has been collected from a part, rather than the whole of the population.
- ✓ Because of the above, sampling errors are restricted to sample surveys only unlike non-sampling errors that can occur in both sample surveys and censuses data.

Sample Size

The "right" sample size for a particular application depends on many factors, including the following:


- Cost considerations (e.g., maximum budget, desire to minimize cost).
- Administrative concerns (e.g., complexity of the design, research deadlines).
- Minimum acceptable level of precision.
- Confidence level.
- Variability within the population or subpopulation (e.g., stratum, cluster) of interest.
- Sampling method.

Required Sample Size[†]
from: **The Research Advisors**


Population Size	Confidence = 95,0%				Confidence = 99,0%			
	Degree of Accuracy/Margin of Error				Degree of Accuracy/Margin of Error			
	0,05	0,035	0,025	0,01	0,05	0,035	0,025	0,01
10	10	10	10	10	10	10	10	10
20	19	20	20	20	19	20	20	20
30	28	29	29	30	29	29	30	30
50	44	47	48	50	47	48	49	50
75	63	69	72	74	67	71	73	75
100	80	89	94	99	87	93	96	99
150	108	126	137	148	122	135	142	149
200	132	160	177	196	154	174	186	198
250	152	190	215	244	182	211	229	246
300	169	217	251	291	207	246	270	295
400	196	265	318	384	250	309	348	391
500	217	306	377	475	285	365	421	485
600	234	340	432	565	315	416	490	579
700	248	370	481	653	341	462	554	672
800	260	396	526	739	363	503	615	763
900	269	419	568	823	382	541	672	854
1 000	278	440	606	906	399	575	727	943
1 200	291	474	674	1067	427	636	827	1119
1 500	306	515	759	1297	460	712	959	1376
2 000	322	563	869	1655	498	808	1141	1785
2 500	333	597	952	1984	524	879	1288	2173
3 500	346	641	1068	2565	558	977	1510	2890
5 000	357	678	1176	3288	586	1066	1734	3842
7 500	365	710	1275	4211	610	1147	1960	5165
10 000	370	727	1332	4899	622	1193	2098	6239
25 000	378	760	1448	6939	646	1285	2399	9972
50 000	381	772	1491	8056	655	1318	2520	12455
75 000	382	776	1506	8514	658	1330	2563	13583
100 000	383	778	1513	8762	659	1336	2585	14227
250 000	384	782	1527	9248	662	1347	2626	15555
500 000	384	783	1532	9423	663	1350	2640	16055
1 000 000	384	783	1534	9512	663	1352	2647	16317
2 500 000	384	784	1536	9567	663	1353	2651	16478
10 000 000	384	784	1536	9594	663	1354	2653	16560
100 000 000	384	784	1537	9603	663	1354	2654	16584
264 000 000	384	784	1537	9603	663	1354	2654	16586

The recommended sample size for a given population size, level of confidence, and margin of error appears in the body of the table.

For example, the recommended sample size for a population of 1,000, a confidence level of 99%, and a margin of error (degree of accuracy) of 3.5% would be 575.

 Change these values to select different levels of confidence.

 Change these values to select different maximum margins of error.

 Change these values to select different (e.g., more precise) population sizes.

[†] Copyright, The Research Advisors (2006). All rights reserved.

The formula used for these calculations was:

$$n = \frac{X^2 * N * P * (1 - P)}{(ME^2 * (N - 1)) + (X^2 * P * (1 - P))}$$

Where :

n = sample size

X^2 = Chi – square for the specified confidence level at 1 degree of freedom

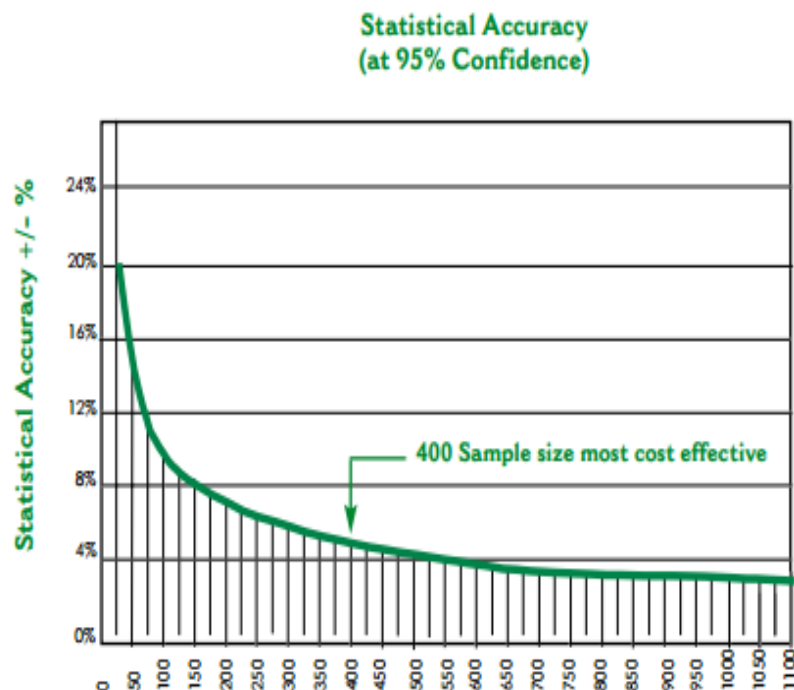
N = Population Size

P = population proportion (.50 in this table)

ME = desired Margin of Error (expressed as a proportion)

Sample Size – What is Magic About the #400?

Statistical accuracy is a function of the sample size. The larger the sample size, the greater the statistical accuracy of the results. Keep in mind that we often look at a subset of the entire sample that we call a “cell” or “cut” (i.e. small companies versus large companies). The statistical accuracy depends on the number of people in each cell. The charts below show that gains in statistical accuracy are not proportional to increases in the sample size. There are “diminishing returns.” At some point, the increase in statistical accuracy may not be worth the additional cost. A sample size of 400 gives a statistical accuracy of $\pm 5\%$ and is often thought of as the most “cost effective” sample size. The table below shows the statistical accuracy calculated for various cell sizes.



Number of Responses per Cell	95% Confidence Level
1000	$\pm 3.1\%$
800	$\pm 3.5\%$
600	$\pm 4.0\%$
400	$\pm 4.9\%$
300	$\pm 5.7\%$
250	$\pm 6.2\%$
200	$\pm 6.9\%$
150	$\pm 8.0\%$

A cell size of 400 with a 95% confidence level can be interpreted as follows: If we repeated the survey 100 times, we would expect the answer to any question to vary less than $\pm 4.9\%$ in 95 of the 100 cases.

**Table I. Minimum effective sample size for countries
(based on the use of a sample of households/address)**

EU-Member States	Households		Persons aged 16 or over to be interviewed	
	Cross-sectional	Longitudinal	Cross-sectional	Longitudinal
Belgium	4 750	3 500	8 750	6 500
Bulgaria	4 500	3 500	10 000	7 500
Czech Republic	4 750	3 500	10 000	7 500
Denmark	4 250	3 250	7 250	5 500
Germany	8 250	6 000	14 500	10 500
Estonia	3 500	2 750	7 750	5 750
Greece	4 750	3 500	10 000	7 250
Spain	6 500	5 000	16 000	12 250
France	7 250	5 500	13 500	10 250
Ireland	3 750	2 750	8 000	6 000
Italy	7 250	5 500	15 500	11 750
Cyprus	3 250	2 500	7 500	5 500
Latvia	3 750	2 750	7 650	5 600
Lithuania	4 000	3 000	9 000	6 750
Luxembourg	3 250	2 500	6 500	5 000
Hungary	4 750	3 500	10 250	7 750
Malta	3 000	2 250	7 000	5 250
Netherlands	5 000	3 750	8 750	6 500
Austria	4 500	3 250	8 750	6 250
Poland	6 000	4 500	15 000	11 250
Portugal	4 500	3 250	10 500	7 500
Romania	5 250	4 000	12 750	9 500
Slovenia	3 750	2 750	9 000	6 750
Slovakia	4 250	3 250	11 000	8 250
Finland	4 000	3 000	6 750	5 000
Sweden	4 500	3 500	7 500	5 750
United Kingdom	7 500	5 750	13 750	10 500
Total of EU Member States	130 750	98 250	272 900	203 850
Iceland	2 250	1 700	3 750	2 800
Norway	3 750	2 750	6 250	4 650
Total including Iceland and Norway	136 750	102 700	282 900	211 300

**Table II. Minimum effective sample size for countries
using a sample or persons**

EU-Member States	Households / Persons aged 16 or over to be interviewed in <u>detail</u>		Persons aged 16 or over to be covered	
	Cross-sectional	Longitudinal	Cross-sectional	Longitudinal
Denmark	5 500	4 250	9 500	7 250
Netherlands	6 500	5 000	11 500	8 750
Slovenia	6 750	5 000	16 250	12 250
Finland	5 000	3 750	8 500	6 250
Sweden	5 750	4 500	9 500	7 500
Non EU countries				
Iceland	3 000	2 000	5 000	3 250
Norway	4 750	3 500	8 000	6 000

Countries using a sample of persons (selected respondents) must select an extra sample to cover population 14 and 15 years old.

Table 4: Certainty plus Minimum coverage percentages, by State (November 2007)

State	State Fixed Sample Size	Certainty Units	Minimum Sample Size	Certainty plus Minimum	% Certainty plus minimum
Alabama	24,090	1,083	14,258	15,341	63.7%
Alaska	4,620	547	3,190	3,737	80.9%
Arizona	16,620	1,660	7,628	9,288	55.9%
Arkansas	15,180	2,025	10,110	12,135	79.9%
California	106,350	8,804	37,673	46,477	43.7%
Colorado	22,740	3,711	11,626	15,337	67.4%
Connecticut	20,670	844	9,311	10,155	49.1%
DC	3,240	306	1,181	1,487	45.9%
Delaware	5,490	734	2,712	3,446	62.8%
Florida	64,020	4,021	29,558	33,579	52.5%
Georgia	28,770	2,273	16,503	18,776	65.3%
Guam	1,800	182	553	735	40.8%
Hawaii	6,210	676	2,506	3,182	51.2%
Idaho	8,160	331	7,208	7,539	92.4%
Illinois	36,840	3,365	16,811	20,176	54.8%
Indiana	31,830	1,763	17,486	19,249	60.5%
Iowa	18,690	854	12,621	13,475	72.1%
Kansas	14,970	850	6,096	6,946	46.4%
Kentucky	19,410	1,070	11,781	12,851	66.2%
Louisiana	22,830	1,145	12,184	13,329	58.4%
Maine	9,240	301	5,337	5,638	61.0%
Maryland	19,830	1,169	8,922	10,091	50.9%
Massachusetts	30,360	1,780	16,690	18,470	60.8%
Michigan	35,550	2,575	19,721	22,296	62.7%
Minnesota	23,190	1,599	9,944	11,543	49.8%
Mississippi	12,480	684	8,569	9,253	74.1%
Missouri	27,000	1,394	12,525	13,919	51.6%
Montana	6,810	615	5,764	6,379	93.7%
Nebraska	10,650	541	6,493	7,034	66.0%
Nevada	10,080	693	4,828	5,521	54.8%
New Hampshire	10,950	335	8,154	8,489	77.5%
New Jersey	35,490	2,289	11,267	13,556	38.2%
New Mexico	9,840	432	6,392	6,824	69.3%
New York	56,880	4,424	19,850	24,274	42.7%
North Carolina	36,780	2,055	20,223	22,278	60.6%
North Dakota	6,330	580	4,907	5,487	86.7%
Ohio	50,820	2,851	21,007	23,858	46.9%
Oklahoma	15,690	965	8,268	9,233	58.8%

State	State Fixed Sample Size	Certainty Units	Minimum Sample Size	Certainty plus Minimum	% Certainty plus minimum
Oregon	19,290	2,666	10,440	13,106	67.9%
Pennsylvania	51,030	2,913	22,398	25,311	49.6%
Puerto Rico	9,810	706	4,677	5,383	54.9%
Rhode Island	6,390	223	1,984	2,207	34.5%
South Carolina	19,920	1,067	13,220	14,287	71.7%
South Dakota	6,240	638	4,538	5,176	82.9%
Tennessee	25,200	1,700	14,172	15,872	63.0%
Texas	76,560	5,905	34,291	40,196	52.5%
Utah	11,430	1,943	7,716	9,659	84.5%
Vermont	5,340	487	3,302	3,789	71.0%
Virgin Islands	1,440	135	609	744	51.7%
Virginia	30,330	1,872	14,719	16,591	54.7%
Washington	25,800	1,380	15,904	17,284	67.0%
West Virginia	11,280	1,126	7,529	8,655	76.7%
Wisconsin	29,760	1,580	18,272	19,852	66.7%
Wyoming	5,040	409	4,238	4,647	92.2%
Totals:	1,215,360	86,276	607,866	694,142	57.1%