13 Collecting Statistical Data

Content:

13.1 The Population ........................................................................................................................................ 2
The Population - what is the population to which the statement applies? ......................................................... 2
The N Value - How many individuals or objects are there in that population? ................................................... 2
Data................................................................................................................................................................. 2
Census .............................................................................................................................................................. 3

13.2 A Survey .................................................................................................................................................. 3
CASE STUDY 2 THE 1936 LITERARY DIGEST POLL FOR PRESIDENTIAL ELECTION ............................................. 4
Convenience Sampling ...................................................................................................................................... 5
Quota Sampling ............................................................................................................................................... 5

13.3 Random Sampling .................................................................................................................................. 6
Simple Random Sampling ............................................................................................................................... 6
Stratified Sampling ......................................................................................................................................... 6
CASE STUDY 4 NATIONAL PUBLIC OPINION POLLS ..................................................................................... 6

13.4 Sampling: Terminology and Key Concepts ............................................................................................. 7
Survey .............................................................................................................................................................. 7
Statistic versus Parameter ............................................................................................................................. 7
Sampling Error ............................................................................................................................................... 7
Chance Error .................................................................................................................................................. 7
Sampling Bias .................................................................................................................................................. 7
Sampling Proportion ....................................................................................................................................... 7

13.99 Themes of Chapter 13............................................................................................................................... 8

SUMMARY:

Example 1. The city has 8325 voters and elections are between 3 candidates. A telephone poll of randomly 680
selected voters was conducted:
a. 306 voters were for the first candidate
b. 272 were for the second candidate
c. 102 were for the third candidate.

Statistical terms: 8325 is a population (N-value is 8325)

\[
\begin{align*}
680 & \text{ is a sample (} n \text{ value is 680)} & \text{Sampling proportion: } & \frac{n}{N} \\
- \text{ census} & \text{ (all members of the population) compared to the survey/poll (a subset of the population)} \\
- \text{ parameter} & \text{ is the value from the population (calculated through the census)} \\
- \text{ statistic} & \text{ is the value from the sample (calculated based on sample and a sampling method)} \\
- \text{ sampling error} & =|\text{parameter} – \text{statistic}| \text{ It is caused by chance error and sampling bias} \\
- \text{ response rate in the survey: } & \text{ response rate } = \frac{\text{number of responses}}{\text{starting } n}
\end{align*}
\]
If the survey has the drop in sample size due to response rate, then the number of responses is used in calculations as a sample size and NOT the initial $n$.

For example: We sent 150 surveys but only 90% came back. Then $150 \times 0.9 = 135$ is the number of the sample size.

Statistic is often represented in percentages: “find the statistic estimating the percentage of the…” – in that case we find a ratio of the specific value over the value for the sample. For example, $\frac{306}{680}$ is the statistic estimating the percentage of voters that voted for the candidate.

In a survey (poll) the method of selecting the population members for the sample is important. The method is called sampling.

In some cases members are pulled from a subset of population and that is called a sampling frame. This happens for practical reasons. Ideally sampling frame would be equal to the target population such was the case in the example of two types of candies in a jar.

**Sampling types:**
- Convenience – ask people for the opinion in the street
- Quota - a specified number of people from each group
- Simple random – randomly select from all students in the school
- Stratified – randomly select certain number of students from each grade in a high school (freshmen, sophomores, juniors, and seniors).

**How to ensure that sample represents the population?**
Problems to avoid: sampling bias - the sample is selected has a wrong sampling frame, self-selection...
non-response bias - when many people do not respond to the survey

---

13.1 The Population

**How to ensure that sample represents the population?**

Problems to avoid: sampling bias - the sample is selected has a wrong sampling frame, self-selection...
non-response bias - when many people do not respond to the survey

**The Population - what is the population to which the statement applies?**

Every statistical statement refers, directly or indirectly, to some group of individuals or objects. In statistical terminology, this collection of individuals or objects is called the population.

The first question we should ask ourselves when trying to make sense of a statistical statement is, “what is the population to which the statement applies?”

**The N Value - How many individuals or objects are there in that population?**

It is important to keep in mind the distinction between the $N$-value– a number specifying the size of the population–and the population itself.

**Data**

The word data is the plural of the Latin word datum, meaning “something given,” and in ordinary usage has a somewhat broader meaning than the one we will give it in this chapter.
For our purposes we will use the word *data as any type of information packaged in numerical form*, and we will adhere to the standard convention that as a noun it can be used both in singular (“the data is...”), and plural (“the data are...”) forms.

**Census**

The process of collecting data by *going through every member of the population* is called a **census**.

Examples: An election day in a city (every single registered voter may vote). Count students in this class. Count candies in the jar. Count number of cars in a parking lot.

The idea behind a census is simple, but it is often not possible to do a census. In these cases the best we can hope for is a good estimate of the *N*-value.

For example in U. S. Census, despite all the efforts it is not possible to get accurate tallies.

**Example 13.4: 2000 Census Undercounts**

”What is the *N*-value of the national population of the United States?”

The 2000 U.S. Census missed counting between 3 and 4 million people.

The reasons: People are constantly on the move. Many distrust the government. In large urban areas many people are homeless or don’t want to be counted.

If the Census undercount were consistent among all segments of the population, the undercount problem could be solved easily. Unfortunately, the modern U.S. Census is plagued by what is known as a **differential undercount**. Ethnic minorities, migrant workers, and the urban poor populations have significantly larger undercount rates than the population at large, and the undercount rates vary significantly within these groups.

Using modern statistical techniques, it is possible to make adjustments to the raw Census figures that correct some of the inaccuracy caused by the differential undercount, but in 1999 the Supreme Court ruled in *Department of Commerce et al. v. United States House of Representatives et al.* that only the raw numbers, and not statistically adjusted numbers, can be used for the purposes of apportionment of Congressional seats among the states.

13.2 **A Survey**

The practical alternative to a census is to **collect data only from some members of the population and use that data to draw conclusions and make inferences about the entire population**.

Statisticians call this approach a **survey** (or a **poll** when the data collection is done by asking questions).

The subgroup chosen to provide the data is called the **sample**, and the act of selecting a sample is called **sampling**.

Ideally, every member of the population should have an opportunity to be chosen as part of the sample, but this is possible only if we have a mechanism to identify each and every member of the population. In many situations this is impossible. Say we want to conduct a public opinion poll before an election. The population for the poll consists of all voters in the upcoming election, but how can we identify who is and is not going to vote ahead of time? We know who the registered voters are, but among this group there are still many nonvoters.

The first important step in a survey is to distinguish the population for which the survey applies (the **target population**) and the actual subset of the population from which the sample will be drawn, called the **sampling frame**.

The ideal scenario is when the sampling frame is the same as the target population—*that would mean that every member of the target population is a candidate for the sample.*

When this is impossible (or impractical), an appropriate sampling frame must be chosen.
Example 13.5: Sampling Frames Can Make a Difference

A CNN/USA Today/Gallup poll conducted right before the November 2, 2004, national election asked the following question: “If the election for Congress were being held today, which party’s candidate would you vote for in your congressional district, the Democratic Party’s candidate or the Republican Party’s candidate?”

When the question was asked of 1866 registered voters nationwide, the results of the poll were 49% for the Democratic Party candidate, 47% for the Republican Party candidate, 4% undecided. When exactly the same question was asked of 1573 likely voters nationwide, the results of the poll were 50% for the Republican Party candidate, 47% for the Democratic Party candidate, 3% undecided.

Clearly, one of the two polls had to be wrong, because in the first poll the Democrats beat out the Republicans, whereas in the second poll it was the other way around. The only significant difference between the two polls was the choice of the sampling frame—in the first poll the sampling frame used was all registered voters, and in the second poll the sampling frame used was all likely voters.

Although neither one faithfully represents the target population of actual voters, using likely voters instead of registered voters for the sampling frame gives much more reliable data. (The second poll predicted very closely the average results of the 2004 congressional races across the nation.)

So, why don’t all pre-election polls use likely voters as a sampling frame instead of registered voters?

The answer is economics. Registered voters are relatively easy to identify—every county registrar can produce an accurate list of registered voters. Not every registered voter votes, though, and it is much harder to identify those who are “likely” to vote. Typically, one has to look at demographic factors (age, ethnicity, etc.) as well as past voting behavior to figure out who is likely to vote and who isn’t. Doing that takes a lot more effort, time, and money.

The basic philosophy behind sampling is simple and well understood—if we have a sample that is “representative” of the entire population, then whatever we want to know about a population can be found out by getting the information from the sample.

If we are to draw reliable data from a sample, we must:

(a) find a sample that is representative of the population
(b) determine how big the sample should be.

These two issues go hand in hand, and we will discuss them next.

Sometimes a very small sample can be used to get reliable information about a population, no matter how large the population is. This is the case when the population is highly homogeneous.

The more heterogeneous a population gets, the more difficult it is to find a representative sample. The difficulties can be well illustrated by taking a look at the history of public opinion polls.

Case Study 2 The 1936 Literary Digest Poll for Presidential Election

The sampling frame for the Literary Digest poll consisted of an enormous list of names that included:

1. every person listed in a telephone directory anywhere in the United States,
2. every person on a magazine subscription list, and
3. every person listed on the roster of a club or professional association.

From this sampling frame a list of about 10 million names was created, and every name on this list was mailed a mock ballot and asked to mark it and return it to the magazine.

Based on the poll results, the Literary Digest predicted a landslide victory for Landon with 57% of the vote, against Roosevelt’s 43%. Amazingly, the election turned out to be a landslide victory for Roosevelt with 62% of the vote, against 38% for Landon.

A large sample does not mean more accurate results.

For the same election, a young pollster named George Gallup was able to predict accurately a victory for Roosevelt using a sample of “only” 50,000 people.
What went wrong with the *Literary Digest* poll and why was Gallup able to do so much better?

When it came to economic status the Literary Digest sample was far from being a representative cross section of the voters because only well off people were included while many at that time struggled due to economic crisis.

When the choice of the sample has a built-in tendency (whether intentional or not) to exclude a particular group or characteristic within the population, we say that a survey suffers from selection bias.

It is obvious that selection bias must be avoided, but it is not always easy to detect it ahead of time. Even the most scrupulous attempts to eliminate selection bias can fall short.

The second serious problem with the *Literary Digest* poll was the issue of nonresponse bias. In a typical survey it is understood that not every individual is willing to respond to the survey request (and in a democracy we cannot force them to do so). Those individuals who do not respond to the survey request are called nonrespondents, and those who do are called respondents.

The percentage of respondents out of the total sample is called the response rate.

Example: For the *Literary Digest* poll, out of a sample of 10 million people who were mailed a mock ballot only about 2.4 million mailed a ballot back, resulting in a 24% response rate.

When the response rate to a survey is low, the survey is said to suffer from nonresponse bias.

The *Literary Digest* story has two morals:

(1) You’ll do better with a well-chosen small sample than with a badly chosen large one, and

(2) watch out for selection bias and nonresponse bias.

Convenience Sampling

In convenience sampling the selection of which individuals are in the sample is dictated by what is easiest or cheapest for the data collector, never mind trying to get a representative sample.

A classic example of convenience sampling is when interviewers set up at a fixed location such as a mall or outside a supermarket and ask passersby to be part of a public opinion poll.

A different type of convenience sampling occurs when the sample is based on self-selection–the sample consists of those individuals who volunteer to be in it.

Convenience sampling is not always bad–at times there is no other choice or the alternatives are so expensive that they have to be ruled out.

We should keep in mind, however, that data collected through convenience sampling are naturally tainted and should always be scrutinized (that’s why we always want to get to the details of how the data were collected). More often than not, convenience sampling gives us data that are too unreliable to be of any scientific and practical value.

With data, as with so many other things, you get what you pay for.

Quota Sampling

Quota sampling is a systematic effort to force the sample to be representative of a given population through the use of quotas–the sample should have so many women, so many men, so many blacks, so many whites, so many people living in urban areas, so many people living in rural areas, and so on. The proportions in each category in the sample should be the same as those in the population.
If we can assume that every important characteristic of the population is taken into account when the quotas are set up, it is reasonable to expect that the sample will be representative of the population and produce reliable data.

The flaw in quota sampling is that, other than meeting the quotas, the interviewers are free to choose whom they interview. This opens the door to selection bias.

### 13.3 Random Sampling

The best alternative to human selection is to let the laws of chance determine the selection of a sample.

**Sampling methods that use randomness as part of their design are known as random sampling methods**, and any sample obtained through random sampling is called a **random sample** (or a probability sample).

#### Simple Random Sampling

The most basic form of random sampling is called **simple random sampling**. It is based on the same principle a lottery is. Any set of numbers of a given size has an equal chance of being chosen as any other set of numbers of that size. In theory, simple random sampling is easy to implement. We put the name of each individual in the population in “a hat,” mix the names well, and then draw as many names as we need for our sample. Of course “a hat” is just a metaphor.

These days, the “hat” is a computer database containing a list of members of the population. A computer program then randomly selects the names.

This is a fine idea for small, compact populations, but a hopeless one when it comes to national surveys and public opinion polls. For most public opinion polls—especially those done on a regular basis the time and money needed to do this are simply not available.

#### Stratified Sampling

The alternative to simple random sampling used nowadays for national surveys and public opinion polls is a sampling method known as stratified sampling.

The basic idea of stratified sampling is to break the sampling frame into categories called **strata**, and then (unlike quota sampling) randomly choose a sample from these strata. The chosen strata are then further divided into categories, called substrata, and a random sample is taken from these substrata.

The selected substrata are further subdivided, a random sample is taken from them, and so on. The process goes on for a predetermined number of steps (usually four or five).

**CASE STUDY 4 NATIONAL PUBLIC OPINION POLLS**

In national public opinion polls the **strata** and **substrata** are defined by a combination of geographic and demographic criteria. For example, the nation is first divided into “size of community” **strata** (big cities, medium cities, small cities, villages, rural areas, etc.).

The strata are then subdivided by geographical region (New England, Middle Atlantic, East Central, etc.). This is the first layer of substrata.

Within each geographical region and within each size of community stratum some communities (called **sampling locations**) are selected by simple random sampling. The selected sampling locations are the only places where interviews will be conducted. Next, each of the selected sampling locations is further subdivided into geographical units called **wards**. This is the second layer of substrata.

Within each sampling location some of the wards are selected using simple random sampling, which are then divided into smaller units, called **precincts** (third layer). Within each ward some of its precincts are selected by simple random sampling, divided into **households** (fourth layer) which are selected by simple random sampling. The interviewers are then given specific instructions as to which households in their assigned area they must conduct interviews in and the order that they must follow.
The efficiency of stratified sampling compared with simple random sampling in terms of cost and time is clear. The members of the sample are clustered in well-defined and easily manageable areas. For a large, heterogeneous nation like the United States, stratified sampling has generally proved to be a reliable way to collect national data.

What about the size of the sample? Surprisingly, it does not have to be very large. Typically, a Gallup poll is based on samples consisting of approximately 1500 individuals, and roughly the same size sample can be used to poll the populations of a small city as the population of the United States. The size of the sample does not have to be proportional to the size of the population.

13.4 Sampling: Terminology and Key Concepts

Survey

As we now know, except for a census, the common way to collect statistical information about a population is by means of a survey.

When the survey consists of asking people their opinion on some issue, we refer to it as a public opinion poll. In a survey, we use a subset of the population, called a sample, as the source of our information, and from this sample, we try to generalize and draw conclusions about the entire population.

Statistic versus Parameter

Statisticians use the term statistic to describe any kind of numerical information drawn from a sample. A statistic is always an estimate for some unknown measure, called a parameter, of the population.

A parameter is the numerical information we would like to have. Calculating a parameter is difficult and often impossible, since the only way to get the exact value for a parameter is to use a census. If we use a sample, then we can get only an estimate for the parameter, and this estimate is called a statistic.

Sampling Error

We will use the term sampling error to describe the difference between a parameter and a statistic used to estimate that parameter.

In other words, the sampling error measures how much the data from a survey differs from the data that would have been obtained if a census had been used.

Sampling error can be attributed to two factors: chance error and sampling bias.

Chance Error

Chance error is the result of the basic fact that a sample, being just a sample, can only give us approximate information about the population.

In fact, different samples are likely to produce different statistics for the same population, even when the samples are chosen in exactly the same way—a phenomenon known as sampling variability. While sampling variability, and thus chance error, are unavoidable, with careful selection of the sample and the right choice of sample size they can be kept to a minimum.

Sampling Bias

Sample bias is the result of choosing a bad sample and is a much more serious problem than chance error. Even with the best intentions, getting a sample that is representative of the entire population can be very difficult and can be affected by many subtle factors. Sample bias is the result. As opposed to chance error, sample bias can be eliminated by using proper methods of sample selection.

Sampling Proportion

Last, we shall make a few comments about the size of the sample, typically denoted by the letter n (to contrast with N, the size of the population).
The ratio $n/N$ is called the **sampling proportion**. A sampling proportion of $x\%$ tells us that the size of the sample is intended to be $x\%$ of the population. Some sampling methods are conducive for choosing the sample so that a given sampling proportion is obtained.

But in many sampling situations it is very difficult to predetermine what the exact sampling proportion is going to be (we would have to know the exact values of both $N$ and $n$). In any case, it is not the sampling proportion that matters but rather the absolute sample size and quality.

Typically, modern public opinion polls use samples of $n$ between 1000 and 1500 to get statistics that have a margin of error of less than 5%, be it for the population of a city, a region, or the entire country.

### 13.99 Themes of Chapter 13

In this chapter we have discussed different methods for collecting data. In principle, the most accurate method is a **census**, a method that relies on collecting data from each member of the population. In most cases, because of considerations of cost and time, a census is an unrealistic strategy. When data are collected from only a subset of the population (called a **sample**), the data collection method is called a **survey**.

The most important rule in designing good surveys is to eliminate or minimize **sample bias**. Today, almost all strategies for collecting data are based on surveys in which the laws of chance are used to determine how the sample is selected, and these methods for collecting data are called **random sampling methods**. Random sampling is the best way known to minimize or eliminate sample bias.

Two of the most common random sampling methods are **simple random sampling** and **stratified sampling**. In some special situations, other more complicated types of random sampling can be used.