Chapter Four

CHOOSING AMONG THE VARIOUS TYPES OF INTERNET SURVEYS

Here, we examine the various types of Internet surveys and the differences among them that factor into deciding what sort of survey is most appropriate for a particular study. At the heart of this decision lies the question of whether a researcher wants to make inferences about some larger population. This chapter deals with the consequences that arise from the answer to that question. (For instance, probability samples generally allow for inferences beyond the sample at hand, whereas convenience samples generally do not.)

Table 4.1 presents the various sampling selection methods related to Internet surveys for the two sampling categories. Later, we discuss each method in some detail.[1]

CONVENIENCE SAMPLING APPROACHES

Convenience sampling is characterized by a nonsystematic approach to recruiting respondents that often allows a potential respondent to self-select into the sample. Any sample in which the probability of a sample member's inclusion in the sample cannot be computed is considered to be a convenience sample. As we noted earlier in this report, convenience samples often require much less time and effort

[1]Other such taxonomies and further discussions on this topic can be found in Couper (2000) and Bradley (1999).

Table 4.1

Sampling Selection Methods for Internet-Based Surveys

Sampling Category	Selection Method
Convenience	Uncontrolled instrument distribution
	Systematic sampling of Web site visitors
	Volunteer panel
Probability	Sample from a closed population list
	Sample from a general population
	Prerecruited panel

to generate than probability samples, and thus are usually less costly. However, statistical inference is much more problematic with convenience samples. For example, in a survey about the environment, respondents who have an active interest in environmental issues may be more likely to self-select into the survey than others. Such a survey would likely overestimate the degree of concern within the general population about the environment.

Nevertheless, convenience sampling can be useful in other ways (as discussed in Chapter Two). It can be extremely valuable for hard-to-reach (although electronically connected) populations. Under certain assumptions, convenience samples can also be used for *model-based inference*.[2] In such a case, it is assumed that the regression model is correctly specified, meaning that all variables that affect the response are included in the model. Generally, a solid theory of how a model should be specified is not available and therefore variable selection procedures are employed. Moreover, it is possible to only disprove, and not prove, such a theory. Therefore, the assumption that the regression model is correctly specified is problematic.

Convenience samples are particularly unsuitable for estimating totals and fractions, which is often desirable in survey sampling.

---

[2]Although model-based inference is often employed in other branches of statistics, it remains controversial among survey statisticians. This may be due to historical developments specific to survey statistics.

Uncontrolled Instrument Distribution

By way of definition, a simple example of an uncontrolled instrument distribution is the posting of a survey on the Web for anyone to fill out. This type of Web survey has become ubiquitous. Certain organizations, including those supplying the daily news, routinely conduct Web polls, ostensibly for the reader's entertainment, and some Web sites exist for no other reason than to host polls (for example, misterpoll.com and survey.net). Participation in these surveys is entirely voluntary and self-selected. Chapter Six contains a case study that illustrates the use of an inexpensive survey with a convenience sample.

Surveys conducted via uncontrolled instrument distribution are "uncontrolled" because anyone with Web access can fill them out, as many times as they desire. There are ways to try to control multiple access by a particular computer user, but savvy users can fairly easily circumvent those safeguards. Similarly, screening questions can be implemented to prevent multiple access by the same individual. Preventing multiple access, however, does not change the fact that the sample constitutes a convenience sample.

In addition, survey sponsors can actively advertise their surveys in various venues in an attempt to encourage survey participation. Web advertising may be used to attract particular types of survey respondents, such as visitors to certain newsgroups or Web sites, just as commercial advertising might be used to attract specific types of consumers. But because the advertised survey cannot be restricted to solely the advertisement recipients, the distribution is still uncontrolled because anyone can have access to it. For an example of a Web survey using advertising, see Schillewaert et al. (1998).

Many uncontrolled instrument distribution surveys are published only on the Web or in newspaper articles. One exception is Coomber (1997), who conducted a survey of drug dealers worldwide. Coomber was interested in the practice of drug dilution (cutting drugs with other substances to increase profits). Specifically, he wanted to find out how common the practice of *dangerous* drug dilution (cutting drugs with substances such as household cleansers) was internationally. Obviously, lists of illegal drug dealers do not exist and therefore Coomber could not construct a sample frame. Instead,

Coomber advertised on newsgroups and directed respondents to a Web survey site. He also sent e-mails to individuals who had posted messages on the newsgroups. (To avoid being subpoenaed to reveal the respondents' e-mail addresses, Coomber did not attempt to learn their identities.) He recommended that respondents access the Web from a public terminal, such as one at a public library, or print the survey out and return it anonymously by postal mail. Coomber received 80 responses from 14 countries on four continents; 40 percent of the responses came from the United States.

Systematic Sampling of Web Site Visitors

Sampling every *nth* person from a sample frame that is ordered in some way is called *systematic sampling*. For instance, it is possible to have surveys "pop up" on the computer screen of every nth visitor to a Web site. One company, Zoomerang (www.zoomerang.com), sells technology that makes it possible to invite only every nth visitor to a site to fill out a survey.

Sampling every nth visitor constitutes a probability sample if one defines the target population as "visitors to this particular Web site." For other target populations, the outcome would be regarded as a convenience sample. In addition, cookies (small pieces of information stored on a Web users' computer) can be used to ensure that Web site visitors are selected to participate in a survey only once (assuming the user's Web browser accepts cookies).

Volunteer Panel

The *volunteer panel* method relies on assembling a group of individuals who have volunteered to participate in future surveys. The individuals are generally recruited into the panel through some form of advertising. Harris Interactive (see Chapter Three) employs a volunteer panel with a database of several million volunteer Web survey participants who were recruited from a variety of sources, including advertising on the Internet. Harris Interactive then conducts surveys using convenience samples drawn from its database.

Harris Interactive believes that generalizable results can be obtained based on convenience samples by using propensity scoring. As noted

in Chapter Three, propensity scoring was invented to deal with selection bias, but has not traditionally been used in the context of surveys. The claim that propensity scoring can successfully adjust for selection bias in volunteer panel surveys is controversial among researchers (see Couper, 2000). Harris Interactive insiders claim to have success with propensity scoring by pointing to accurate predictions of election outcomes (Taylor, 2000).

Berrens et al. (2001) compared an RDD survey with identical surveys conducted by Harris Interactive and Knowledge Networks. Despite the large sample sizes, Berrens et al. found that when demographic variables (including income) are adjusted for via regression, all three surveys yielded statistically indistinguishable results on several questions. On the other hand, in a matched comparison study of results from a conventional RDD survey, a Knowledge Networks survey, and a Harris Interactive survey, Chang (2001) found significantly different results among the three methods. In Chapter Six, we present a case study on a Harris Interactive survey.

PROBABILITY SAMPLING APPROACHES

If a probability sample is desired, how to go about obtaining a sample frame that covers most or all of the target population becomes a crucial issue. The nature of the target population is relevant to our discussion here. We distinguish between closed target populations and open, or general target, populations.

Sampling from a Closed Population

We refer to target populations within organizations that maintain some sort of list of their membership as *closed populations* (for example, lists of company employees, university staff members, or magazine subscribers). It is usually fairly easy to construct sample frames for these groups. Even if an organization does not maintain a directory of its members' e-mail addresses (as in the case of the U. S. Air Force, which is discussed in Chapter Six), there may still be a systematic way of constructing those addresses (for example, firstname.lastname@airforcebase.mil). Or, it might be possible to reach individuals via regular internal company mail. In short, there is

usually an obvious way to construct a sample frame, which then makes it feasible to draw a probability sample.

Sampling from General Populations

In this report, we refer to populations other than closed populations as "general populations" (for example, residents of California or patients who have reported adverse drug reactions). Members of general populations are more difficult to contact because a list of e-mail addresses with a wide enough coverage to serve as the sample frame is not usually available. In addition, for the Internet, non–list-based sampling alternatives are not available.[3]

Although e-mail lists with wide coverage are not currently available, that situation may change in the future. Right now, the only way to recruit a probability sample is by contacting potential respondents through some conventional means (generally, by mail or phone). The respondents can then be asked to respond to a survey via the Web (or by another mode or set of modes). The problem with this option is that the cost savings that can be realized through an entirely Internet-based survey process are greatly reduced.

---

[3]List-based sampling approaches require enumeration of an entire population (such as by e-mail address). There are non–list-based alternatives, however. For example, RDD does not require an enumeration of the population, and there are other less-popular methods (for example, area sampling). However, no equivalent to RDD or another similar method exists with the Internet. If such an alternative could be developed, it would mean sending large numbers of unsolicited e-mails. This approach, however, would likely face resistance from Internet service providers and from those advocating against "spam" (junk e-mail), and there would be legal challenges in some U.S. states. In fact, the unsolicited mass distribution of spam *may* be illegal. (Note that RDD is unsolicited phone calling, which is *not* illegal). According to U.S. Code Title 47, Section 227(a)(2)(B), a computer/modem/printer meets the definition of a telephone fax machine and according to Section 227(b)(1)(C), it is unlawful to send any unsolicited advertisements to such equipment. In addition, according to Section 227(b)(3)(C), a violation of this law is punishable by action to recover actual monetary loss, or $500, whichever is greater, for each violation. Whether a computer meets the definition of a fax machine and whether this portion of the U.S. Code actually applies to e-mail spam are controversial matters and apparently have not been tested in court. However, even if spam is legal, there is significant resistance to it within the Internet community to the extent that, once identified, "

" are often denied service by Internet service providers.

If an Internet-based response mode is used, potential respondents must first be contacted through a conventional mode and either directed to a Web site or their e-mail address must be collected for subsequent distribution of an e-mail survey instrument. Given the as-yet-incomplete penetration of the Internet to the general population, this approach currently implies that (1) mixed modes must be used for response so that potential respondents without Internet access can respond; *or* (2) those without Internet access must be provided with the requisite hardware and software as part of the survey effort;[4] *or* (3) researchers must be willing to accept a considerable discrepancy between the sample frame and the target population. Chapter Six contains a case study of a survey in which a general population was contacted via postal mail and then asked to respond via the Web.

Prerecruited Panel

A *prerecruited panel* is a group of potential survey respondents, recruited by some probabilistic method, who are available for repeated surveying. A good example of a firm that uses prerecruited panels is Knowledge Networks, which recruits a panel of individuals via RDD to participate in ongoing surveys. Panelists receive three or four surveys a month requiring between 10 and 15 minutes each to complete. Sampling is controlled such that panelists are not given more than one survey on a given topic in a three-month period.

With both volunteer and recruited panels, one concern that researchers have is that participants may tire of filling out surveys, a condition called "panel fatigue," or learn to provide the easiest responses, a phenomenon called "panel conditioning." There is evidence to support that panel conditioning does happen: Comparing a Web survey conducted by Knowledge Networks and an RDD survey, each using identical questionnaires, Berrens et al. (2001) reported that panel participants gave a considerably higher percentage of "don't know" responses than panelists in the RDD survey. An alternative explanation for the higher rate of "don't know" responses on the Web could be due to the survey mode and design of the

[4]For cost reasons, this approach makes sense only for a panel in which respondents can be used again for other surveys.

instruments rather than panel conditioning. Whereas Web surveys typically offer an explicit "don't know," in telephone surveys, "don't know" responses are usually not offered and are often probed when used by the respondent.

A HYBRID SAMPLING APPROACH: COMBINING A CONVENIENCE SAMPLE WITH A PROBABILITY SAMPLE

Because it can be relatively inexpensive to obtain a convenience sample from the Web, it is reasonable to ask whether there are advantages to combining a large convenience sample with a probability sample. The hope is that the resulting larger combined sample might be more precise than the random sample, or that the probability sample can be used to correct any bias in the convenience sample, again resulting in a larger sample and a more precise result. We have investigated this possibility and the details are given in Appendix C.

We found that it is futile to attempt to adjust the convenience sample because it provides no additional information for any subsequent estimation. It is also not useful to combine an unadjusted convenience sample with a probability sample unless the bias from the convenience sample is known to be very small and the probability sample has at least several thousand respondents. Furthermore, in most, if not all, circumstances, there is no way of knowing the magnitude of the bias in advance. Thus, the addition of a convenience sample to a probability sample is not useful in practice.

SUMMARY

This chapter has focused on the most crucial consideration that researchers need to make before conducting a survey: whether they require a convenience sample or a probability sample. Choosing a probability sample has implications in terms of how respondents can be contacted—for instance, advertising on-line or in newspapers is not an option. Except for closed populations with well-defined e-mail address lists or a standardized nomenclature, if the research requires a probability sample, a conventional contact mode (such as RDD) must be used. If a convenience sample will suffice, however, the survey may be conducted entirely electronically.