Target Populations, Sampling Frames, and Coverage Error

Professor Ron Fricker
Naval Postgraduate School
Monterey, California

Goals for this Lecture

- Define survey sampling terms, including
  - Target population, survey population, sampling frame, element
  - Coverage, undercoverage, ineligible units
- Discuss frame coverage issues and some solutions
- Discuss sampling issues related to web-based and e-mail-based surveys
  - When can an “all electronic” survey approach work and when not?
  - What are the (current) difficulties with using these survey modes for general populations?

Terminology (1)

- The target population is the group of elements to which the researcher wants to make inference
  - At least theoretically, the population is finite and can be counted
- The fundamental units of the population are elements
  - Often, elements are persons
  - They can also be households, housing units, parts of an organization, etc.

Terminology (2)

- The survey population is a subset of the target population (often resulting from practical survey considerations)
  - Using RDD, the survey population is all US households with a landline
  - In CES, employers have to be in business for several months
- The sampling frame is used to identify the elements of the population
  - Via explicit or implicit enumeration

Coverage Properties of Sampling Frames

- A target population element that is in the sampling frame is covered
- Undercoverage is the fraction of the total population not covered by the sampling frame
- Ineligible units are those elements in the sampling frame that are not part of the total population

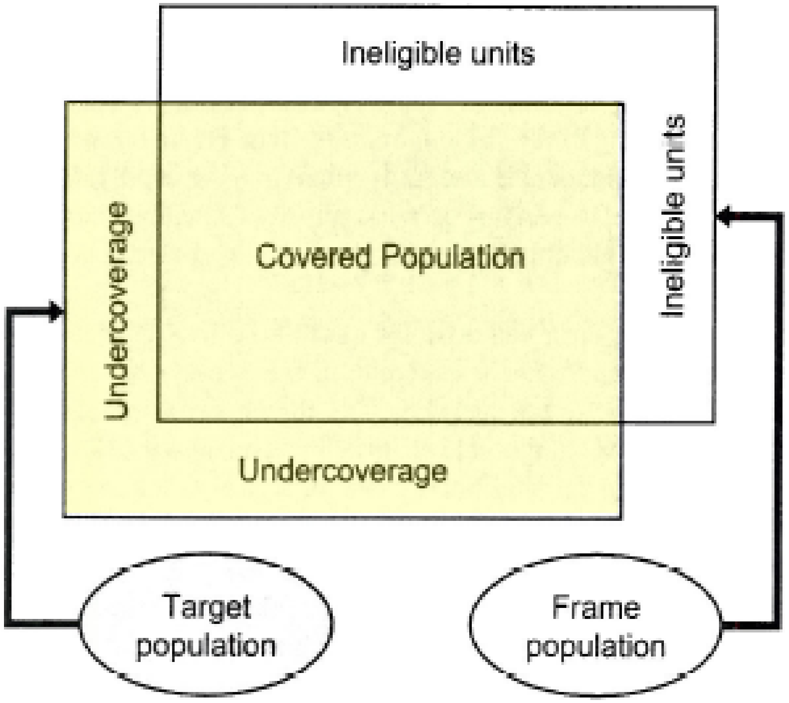


Figure 2.6 Coverage of a target population by a frame.

Sampling Frame Issues

- A sampling frame is perfect if there is a one-to-one mapping from frame to population elements
  - Duplication occurs when multiple frame elements map to one total population element
  - Clustering occurs when multiple total population elements map to one frame element

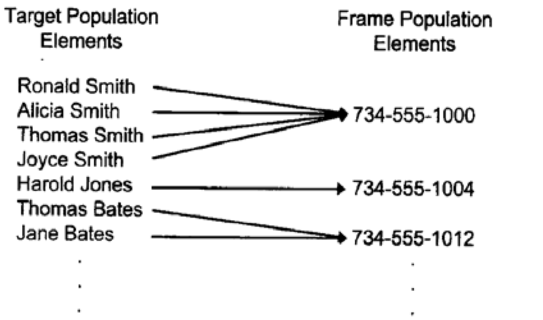


Figure 3.1 Cluster of target population elements associated with one sampling frame element.

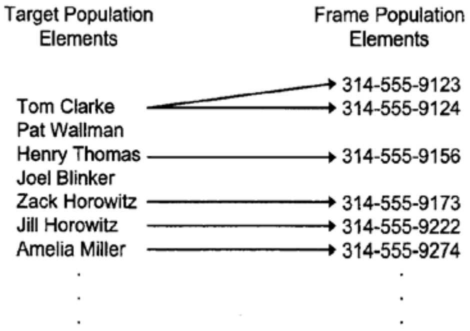


Figure 3.2 Duplication of target population elements by more than one sampling frame element.

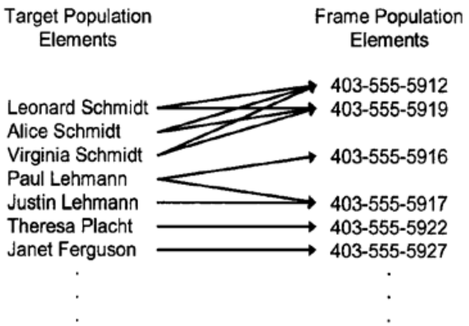


Figure 3.3 Clustering and duplication of target population elements relative to sampling frame elements.

Frame Issues for Household and Person-level Surveys (1)

- Phone number-based frames for sampling households and/or persons
  - Issues:
    - Undercoverage of households without landlines
      - A growing problem with proliferation of cell phones
    - Overcoverage for households with more than one landline
    - Duplication when using lists if same number listed under multiple names
      - With electronic lists, can sort by phone number and remove duplicate phone numbers
      - Can also manage problem during sampling
    - Screening required for non-residential phone numbers

Frame Issues for Household and Person-level Surveys (2)

- Area frames for sampling households and/or persons
  - Area frames require multi-stage sampling
    - First sample areas
    - Then make listing of addresses in sampled areas
    - Then enumerate and sample individuals
  - Issues:
    - Undercoverage if list of addresses within area incomplete
    - Duplication if people have more than one residence
    - Clustering, since housing units contain multiple people

Frame Issues for Household and Person-level Surveys (3)

- Web or e-mail surveys for households and/or persons
  - – Issues:
    - No global list (frame) exists for general population
      - – Sampling frames for e-mail only exist for specific organizations
      - – No practical equivalent of RDD exists
    - Most survey organizations classify sending unsolicited e-mail as unethical (i.e., spam)
- More at the end of the lecture…

Frame Issues for Customer, Employee, Other Members of Organization Surveys

- Studies of customers, employees, members of organizations tend to use list frames
  - E-mail contact/surveys often feasible for these groups
  - Undercoverage can be an issue if lists are out-of-date
    - Also, depending on how list generated, can miss portions of population (e.g., payroll-based list misses volunteers)
    - Inclusion of ineligible units can also be an issue
  - Duplication can be an issue, but likely rare for employee, organization-based lists
    - For customer surveys, if list based on transactions, care must be given to duplicates and what resulting frame actually represents

Frame Issues for Organizational Surveys

- Organizational populations very diverse
  - Sampling frames are often lists of units
- For business units, significant variation in size can be an issue
  - May need to stratify on size and oversample
- In commercial world, population is likely to be highly dynamic
  - Business come and go, merge, etc.
  - Frame may need constant updating

Frame Issues for Event-based Surveys

- Surveys may target events
  - Periods of deployment, enrollment at NPS, etc.
- Often begin with a frame of persons and then screen for event
  - Some persons may have experienced multiple events, so clustering can occur
- Can also use frame of time units
  - I.e., sample customers exiting a store at certain times (of the day, of the week, of the year, etc)
  - Time use surveys sample at random points in time

Reducing Undercoverage

- There are remedies for reducing sampling frame problems
  - But they do not always eliminate undercoverage
- Also, note that what is relevant is how undercoverage affects the sample statistics
  - For some it may be negligible and others significant
- Can represent coverage bias as $\bar{Y}_C - \bar{Y} = \frac{U}{N}\left(\bar{Y}_C - \bar{Y}_U\right)$
  - Note it’s a function of both undercoverage amount and difference between the means

Strategy: Multiple Frame Design

- Idea: Supplement the principle frame with one or more auxiliary frames
  - E.g., supplement RDD with area sampling to get at those without a landline phone
- Must adjust results to account for those that can be selected in both frames
- Multiple frames may require multiple modes
  - Much research still required for problem of multiple frames combined with multiple modes

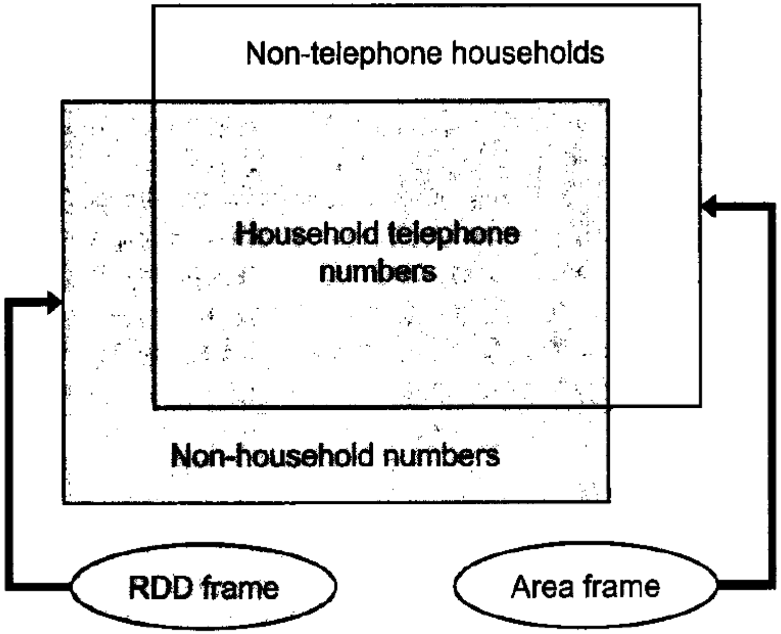


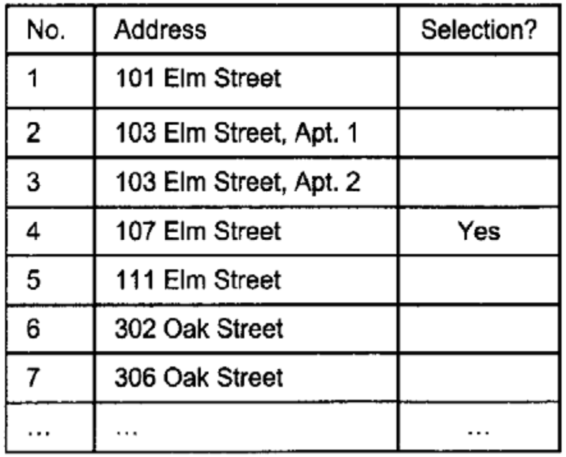
Figure 3.6 Dual frame sample design.

Strategy: Broaden Definition of Who Can Be in Sampling Frame

- Idea: Broaden the definition of who can be in the sampling frame, then screen out as necessary
  - Incurs extra costs and requires additional effort
- Example: Surveys of housing units, where an individual is asked who lives in the house, frequently misses some people
  - Rather than ask “Who lives here?” ask broader questions about who slept or ate in the unit the previous day, who has a key, who receives mail there, etc.
  - Then use other questions to screen out as necessary
- Issue: Since these questions have to come first, can increase number of survey refusals / nonresponse

Strategy: The Half-Open Interval

- Idea: Supplement frame with information gained during selection process
  - Useful for lists that may be missing some entries
- Example: When sampling households, the sampling unit is not the address, but all addresses from the one selected up to the next one on the list



No.	Address	Selection?
1	101 Elm Street	
2	103 Elm Street, Apt. 1	
3	103 Elm Street, Apt. 2	
4	107 Elm Street	Yes
5	111 Elm Street	
6	302 Oak Street	
7	306 Oak Street	
...	...	...

Figure 3.4 Address list for area household survey block.

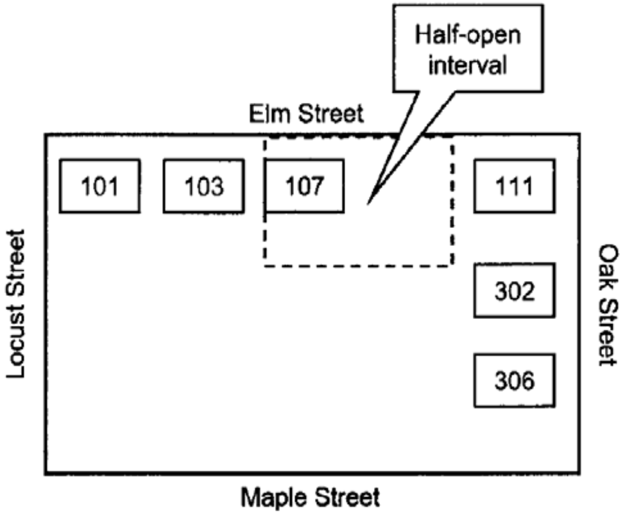


Figure 3.5 Sketch map for area household survey block.

Strategy: Multiplicity Sampling

- Idea: Each unit selected into the sample is asked about members of a well-defined network
  - E.g., An individual is contacted via RDD and asked questions about his/her adult siblings
    - Issue: Individuals may have multiple chances of being selected
    - Need to appropriately weight to adjust
- Similar idea: snowball sampling
  - Locate additional potential survey respondents via current respondents
  - Useful for rare conditions where sampling of general population would be inefficient
  - Non-probability method, so inference difficult/impossible

Doesn’t the Web Solve Everything?

- Today anyone can do a web survey:

SurveyMonkey.com
because knowledge is everything
KeySurvey
Survey software everyone can use.™
SuperSurvey™
Change The Way You Ask Questions
zoomerang™
Easiest Way to Ask, Fastest Way to Know™
SurveyCrafter
Survey Said™ Survey Software
Cool Surveys
1
4
FREE! Polling Tools and Survey Tools!

Types of Internet-based Surveys

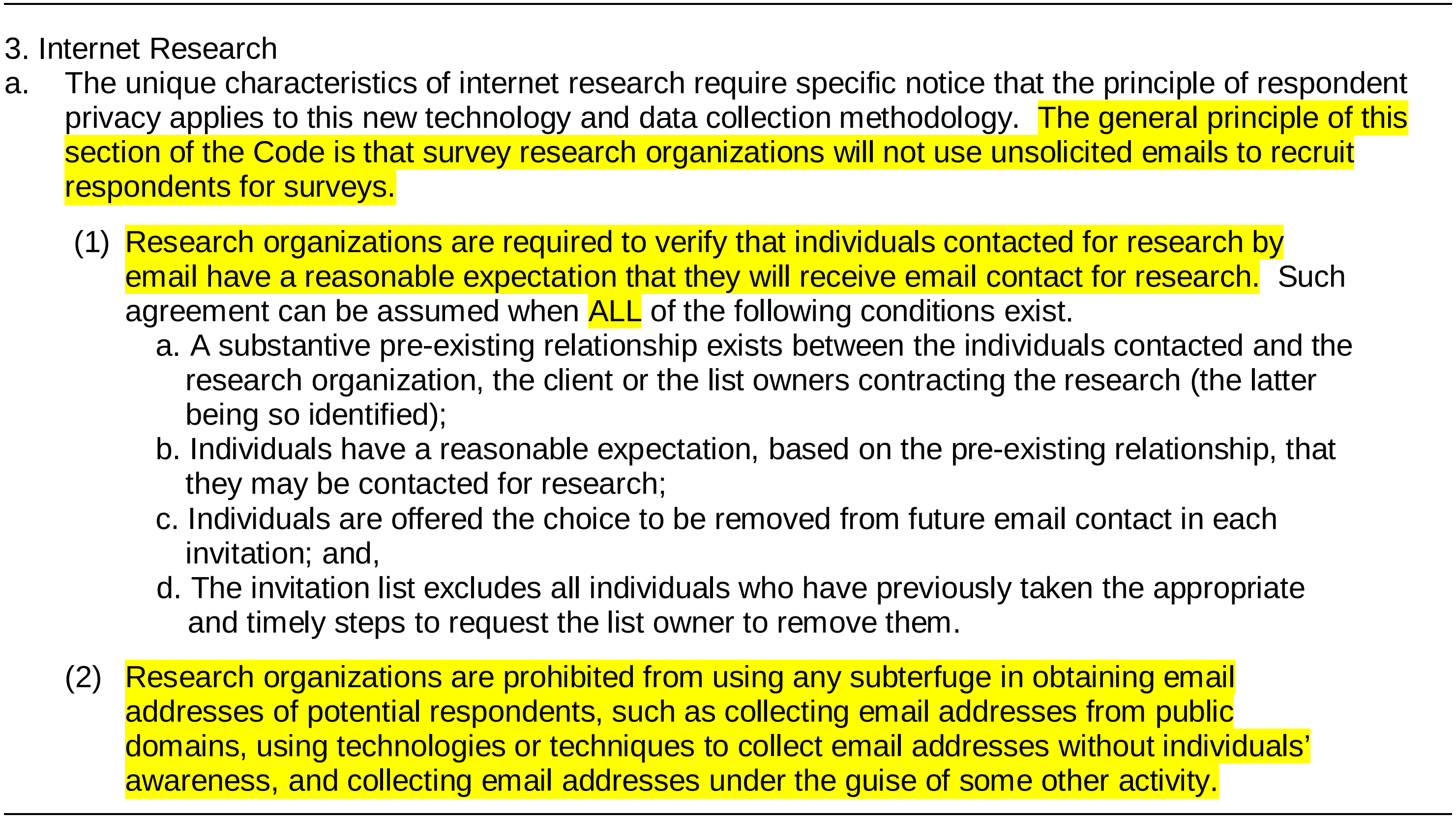
Sampling Method	Web	E-mail
Probability-based		
Surveys using a list-based sampling frame	✓	✓
Surveys using non-list-based random sampling	✓	✓
Intercept (pop-up) surveys	✓	
Mixed mode surveys with Internet-based option	✓	✓
Pre-recruited panel surveys	✓	✓
Non-probability		
Entertainment polls	✓	
Unrestricted self-selected surveys	✓	
Surveys using ‘Harvested’ E-mail Lists (and Data)	✓	✓
Surveys using volunteer (opt-in) panels	✓	

Table 2. *Types of Internet-based surveys and associated sampling methods*.

Internet Surveys Pose Significant Challenges for General Populations

- Internet users do not reflect the general population
  - Frame and coverage bias
- Very hard to impossible to generate a sampling frame
  - There is no master list of e-mail addresses
- Recruiting respondents via the web can introduce bias
- Unethical/illegal to solicit survey participation via e-mail (spam)

Unsolicited E-mail is Spam



3. Internet Research

a. The unique characteristics of internet research require specific notice that the principle of respondent privacy applies to this new technology and data collection methodology. The general principle of this section of the Code is that survey research organizations will not use unsolicited emails to recruit respondents for surveys.

(1) Research organizations are required to verify that individuals contacted for research by email have a reasonable expectation that they will receive email contact for research. Such agreement can be assumed when ALL of the following conditions exist.

a. A substantive pre-existing relationship exists between the individuals contacted and the research organization, the client or the list owners contracting the research (the latter being so identified);

b. Individuals have a reasonable expectation, based on the pre-existing relationship, that they may be contacted for research;

c. Individuals are offered the choice to be removed from future email contact in each invitation; and,

d. The invitation list excludes all individuals who have previously taken the appropriate and timely steps to request the list owner to remove them.

(2) Research organizations are prohibited from using any subterfuge in obtaining email addresses of potential respondents, such as collecting email addresses from public domains, using technologies or techniques to collect email addresses without individuals' awareness, and collecting email addresses under the guise of some other activity.

Figure 2. *Excerpt from Council of American Survey Research Organizations Code of Standards and Ethics for Survey Research, accessed online at www.casro.org/codeofstandards.cfm on September 30, 2006. Highlighting added.*

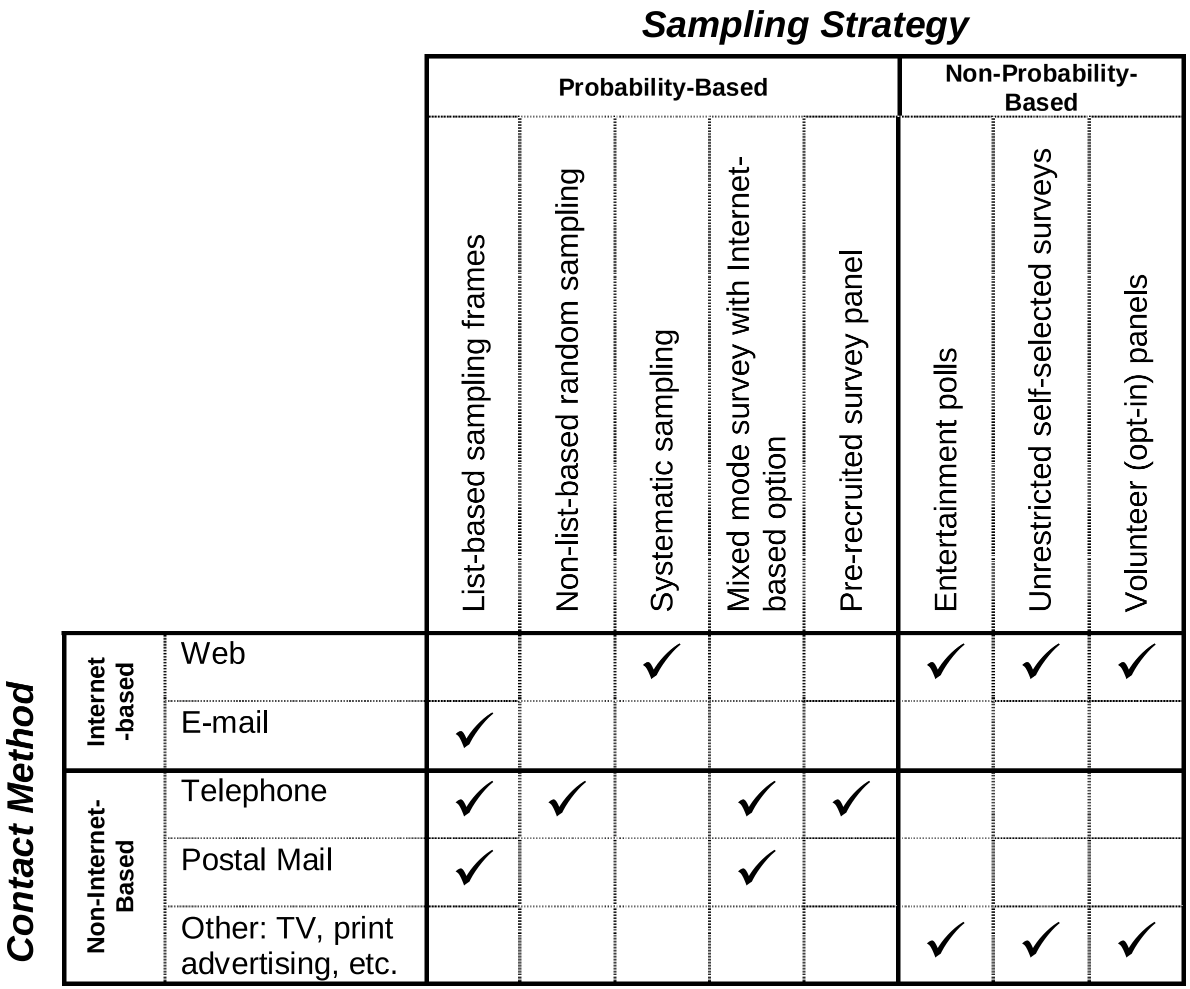
Problems Mitigated for Some Populations

- Internet-based surveys within individual organizations often easier
  - List of e-mail addresses or common e-mail address syntax often available
  - No problem with e-mail as spam
- But issues can remain
  - How complete is the e-mail address list?
  - Do all respondents have access to the Internet/ e-mail?
  - Non-response issues and bias

Bigger Samples Not Always Better

- When conducting Internet-based surveys, temptation is to (attempt to) survey *everyone* Why not? It’s cheap and easy…
- Example: Survey2000
  - Goal: Quantify how often people have moved, what role they play in their communities, and how geography has shaped their tastes in food, music, and literature
- More than 80,000 surveys initiated and 50,000+ completed

Summary



Sampling Strategy

			Probability-Based					Non-Probability-Based		
			List-based sampling frames	Non-list-based random sampling	Systematic sampling	Mixed mode survey with Internet-based option	Pre-recruited survey panel	Entertainment polls	Unrestricted self-selected surveys	Volunteer (opt-in) panels
Contact Method	**Internet -based**	Web			✓			✓	✓	✓
		E-mail	✓							
	Non-Internet-Based	Telephone	✓	✓		✓	✓			
		Postal Mail	✓			✓				
		Other: TV, print advertising, etc.						✓	✓	✓

Table 3. *Sampling strategies* for Internet-based surveys *by contact mode.*

Entirely Web-Based Surveys Generally Restricted to Non-Probability Samples

- Exception is systematic sampling for pop-up/intercept surveys
  - Predominantly use is customer satisfaction surveys for websites or web pages
- Respondent contact can also be conducted via traditional (non-Internet-based) media

Inferential Research Possibilities Constrained with Internet Contact Mode

- E-mail is only useful as a contact mode if a list of e-mail addresses is available
  - List is an actual or *de facto* sampling frame
- Population of inference usually quite limited when using an e-mail address list sampling frame
  - It is generally the sampling frame itself
- Attempted census of entire e-mail list may limit the survey results
  - Nonresponse and other biases may preclude generalizing even to the sample frame

Entirely Internet-Based Surveys Generally Unsuitable For Inferential Research

- If all members of the population of inference do not have e-mail/web access, then contact mode will *have* to be a non-Internet-based medium
- Survey will also have to be conducted using a mixed-mode so that those without Internet access can participate
  - Else, lack of a non-Internet-based survey mode will result in coverage error with the likely consequence of systematic bias
- Pre-recruited panels can provide ready access to pools of Internet-based survey respondents
  - But to allow generalization to general population, panel needs to be recruited using probability sampling methods from the general population
  - And, even so, need to carefully assess whether panel is likely to be subject to other types of bias

What We Have Covered

- Defined survey sampling terms, including
  - Target population, survey population, sampling frame, element
  - Coverage, undercoverage, ineligible units
- Discussed frame coverage issues and some solutions
- Discussed sampling issues related to web-based and e-mail-based surveys
  - When can an “all electronic” survey approach work and when not?
  - What are the (current) difficulties with using these survey modes for general populations?