THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

ASA
AMERICAN STATISTICAL ASSOCIATION

Surveys and Polls: What contributes to quality and cost?

Michael D. Larsen,
The George Washington University
Department of Statistics
Dirksen Building, Room SD – 562
10/10/12; 2-3:30pm

Outline

- In the news
- Steps in the survey process
- Sampling variability
- Non sampling errors
- Technology and changes in society

First, a brief word from the American Statistical Association

American Statistical Association

- ASA is 2nd oldest professional association with more than 18,000 members. http://www.amstat.org
- Statistics is science of collecting, analyzing and understanding data, and so permeates all sciences
- The ASA is offering these classes because we believe statistical critical thinking can be very helpful to staffers as they confront data, studies, reports, … of all sorts
- Please provide us feedback on today's class and suggest future class topics: Steve Pierson, spierson@amstat.org

In the News

Survey and census results are reported every week in the news.

Surveys are conducted by numerous federal (http://www.fedstats.gov/), state, and local agencies and by a multitude of organizations.

Many federal programs require data collection that is mandated by Congress.

In the news - Economic

- Unemployment, employment, labor force – Current Population Survey (CPS) - Census/BLS
- Income and earnings, health insurance, poverty - American Community Survey (ACS) - Census
- Cash and noncash income, government transfer programs - Survey of Income and Program Participation (SIPP) - Census
- Characteristics of college graduates and those with degrees in science and engineering – National Survey of College Graduates (NSCG), Survey of Doctorate Recipients (SDR) - NSF

In the news - Environment

- Land use and natural resource conditions and trends – erosion, wetlands, prime farm land - National Resources Inventory (NRI) - USDA
- Agriculture – Census of Agriculture, Agricultural Resource Management Survey (ARMS) - USDA
- Forests: status, trends, health, ownership, sustainability, production – Forestry Inventory and Analysis (FIA) Program
- Also oceans, Chesapeake Bay, lakes, rivers, fishing, etc.

In the news – Health and Society

- Health – National Health Interview Survey (NHIS), National Health and Nutrition Examination Survey (NHANES), National Immunization Survey , BRFSS; NCHS/CDC.
- Mental health – National Survey on Drug Use and Health; Substance Abuse and Mental Health Services Administration (SAMHSA)
- Bureau of Justice Statistics: courts, prisons, crimes, etc.; National Crime Victimization Survey (NCVS)
- Bureau of Transportation Statistics: commuting, airlines, border crossings, commodity flows, ….
- Surveys of military and their families: DOD, Defense Manpower Data Center (DMDC)

In the news – and of course …

- Election polls – Gallup and many, many others
- Marketing – ubiquitous
- Advocacy – many organizations send surveys to members or as part of fund raising.

Survey Components

First, lets look at the research process. Then lets think about how a survey furthers our search for answers.

Research Process

- Begin with a question: critical gap in knowledge.
- Do you need data to answer the question?
  - Clarify concepts ** non trivial
  - Define variables that be measured ** non trivial
  - Decide how to measure ** non trivial
- Gather data.
  - Experiment – e.g., randomized controlled clinical trial
  - Observational study – e.g., retrospective study on health
  - Census – collect information for the entire population
  - Survey – collect information on a subset of the population

Research process

If you plan to use a survey to gather data for investigation of your question,

- Plan the survey; anticipate biases.
- Collect data; process/check the data.
- Estimation and inference; answer the question.

Did the effort answer your question? Have new questions been identified? Repeat the process.

Survey Components

- Population – what group do you want to study? Adults in the U.S. Suppose we want to ask about financial knowledge, planning, and related topics.
- Target population - more concrete than general population. Adults in households (with telephones?) in the U.S. (with some exclusions)
- Frame or list - actually usable for contacting
  - Units (single elements or groups of elements). The entire household , or a particular adult (most responsible finance)?
  - Explicit list. Census MAF, Postal DSF, other address listing
  - Implicit list. Random digit dialing (RDD) telephone survey
  - Area frame. ¼ section of land (environmental)
  - Also, adaptive designs are used in some difficult situations. Social networks, environmental (animals/birds).

Survey Components, part 2

- Contact mode: in-person (paper/pencil; computer assisted), mail, phone, Internet, or multiple modes.
  - Send to the address a pre-notification postcard followed by a paper survey form with a return envelope.
- Response mode: same or different from contact mode.
  - In the postcard we could offer a web response option.
- Sampling scheme – how do you select the units that are included in the survey? How many units?
  - Suppose we control sampling so that every state is represented and areas with higher unemployment and poverty are oversampled relative to population density.

Survey components, part 3

- Questionnaire development
  - Develop questions or measurements methods: experts, focus groups, etc.
  - Construct questionnaire or other data collection form (put all the questions and information together)
  - Pre-test and revise questionnaire/form
- Data collection
  - Train interviewers, data collectors; standardize process for quality
  - Plan how quality will be monitored and evaluated.
  - Do you offer an incentive for participation? Is participation mandatory by law?
- Data entry and preparation
  - Enter data (if paper from) and check for accuracy
  - Edit and code data (e.g., capture comments in systematic way; find typos and other illogical entries)
  - Edit data file for consistency; compute new variables

Survey Components, part 4

- Analysis/estimation strategy, including survey weighting, variance estimation
  - Survey weights adjust for differential representation of parts of the population in the survey
  - Estimates, such as the percent of the adult population that is very or somewhat knowledgeable about financial planning, have associated uncertainty due to sampling (standard error, margin of error, variance)
  - Exploratory data analysis to find interesting patterns
- Plans for dealing with problems (e.g., non response)
  - Will you sample extra people? Imputation (filling-in) missing values? Survey weighting? Statistical modeling?

Sampling Scheme

How do you select the units that are included in the survey? How many units?

Sampling scheme

- Why sample? Impractical or impossible to observe an entire population.
- Goal: Select a sample that is representative of the entire population. Use data from the sample to estimate characteristics of the population.
- For scientific surveys, probability sampling is preferred.

Probability sampling

- A probability sample is a sample in which each unit in the population has a known, nonzero probability of being included in the sample
- Positive probability
  - Every unit has a chance of being included in the sample
  - No portion of the population is omitted
- Probability of inclusion is known by researchers
  - We can quantify the probability of a unit of being included in the sample

Non probability sampling

- Various names
  - Convenience sample
  - Volunteer sample
  - Quota sample
  - Purposive sample
- Key feature: participants choose themselves to some degree; researcher does not have full control; researchers does not know propensity to participate for everyone in the population
- Selection bias: those who volunteer are different on average in some important way from those who do not
- Information can still be useful in some circumstances

Simple random sampling

- Population size *N*; Sample size *n*
- All samples of size *n* have equal probability
- Imagine putting *N* tokens into a hat and picking *n* randomly; this should be a fair process; it can be done with a computer
- Probability that any unit is included in the sample is *n*/*N*.
- Survey weight for each unit in the sample is *N*/*n*

Simple random sampling

Unbiased: on average, the arithmetic mean in the sample is equals the arithmetic mean in the population

Accuracy:

- The variance of the sample mean is $\frac{\sigma^2}{n}(1-\frac{n}{N})$
- The standard error (SE) of the sample mean is the square root of the variance.
- The SE is estimated using the sample standard deviation (s): $s\sqrt{\frac{1}{n}(1-\frac{n}{N})}$ or $\sqrt{\frac{p(1-p)}{n}\left(1-\frac{n}{N}\right)}$
- The margin of error of the sample mean is a multiplier (1.96 or about 2) times the SE

Example: use Census of Governments as a frame

The U.S. Census Bureau conducts a Census of Governments of all state and local government organization units every 5 years, for years ending in 2 and 7, as required by law under Title 13, U.S. Code, Section 161. The collection of these data will

- Identify the scope and nature of the nation's state and local governments.
- Classify local government organizations, powers, and activities.
- Provide authoritative benchmark data on public finance and public employment.
- Measure state and local governments' fiscal relationships.

Example of SRS

- Government units in U.S. (2012): 89,004
- General purpose units: 38,917
- County units: 3,031

- Lets imagine taking a simple random sample of size ***n*=101** from each of these three groups. How will estimated standard errors of the mean compare in these three populations?
- Suppose 40/101 of the sample governments have an auditing procedure of a particular kind in place.

Example of SRS: SE in 3 populations

- Government units in U.S. (2012): 89,004
- General purpose: 38,917
- County: 3,031

Virtually tied!

$$0.04872 = \sqrt{\frac{0.4(1-0.4)}{101}\left(1-\frac{101}{89004}\right)} \qquad \sqrt{\frac{0.4(1-0.4)}{101}\left(1-\frac{101}{N}\right)}$$

$$0.04868 = \sqrt{\frac{0.4(1-0.4)}{101}\left(1-\frac{101}{38917}\right)}$$

$$0.04793 = \sqrt{\frac{0.4(1-0.4)}{101}\left(1-\frac{101}{3031}\right)}$$

Example: summary statement

If 40% of the sample of counties responded that they have the audit system in place, then one could say that with 95% confidence, the proportion of county government units nationwide that have the audit system in place is between 31% and 49%.

- Standard error: 0.048
- Margin of error: 0.048*1.96 = 0.09
- 40%-9% = 31%; 40%+9% = 49%.
- For all government units, the standard error is a little bigger and the 95% confidence interval would be 30% to 50%.
- To decrease standard error and make a more precise statement, one could increase sample size.

Survey errors

Total survey error is caused by sampling and non sampling errors. Sampling errors arise due to sampling and not taking a census. Non sampling errors arise due to problems of various kinds.

Problems/sources of bias in sample surveys

- Selection bias - not using a probability sampling scheme; using a convenience or other non probability sample
- Frame deficiency - population coverage issues – over coverage or under coverage
- Data collection errors – collect from wrong person or don't follow directions
- Data processing errors – typographical errors and others
- Non response – some sampled units do not participate or fail to provide some information (e.g., income, illegal drug use)
- Response errors – respondents make mistakes (recall error related to timing of event) or do not give accurate answers (underestimate unhealthy foods)
- Specification error – questions do not relate exactly to concept being studied (how do you define unemployment?)

What could be some limitations of the survey of governmental units?

- Frame: some government units not listed and some government units no longer in existence.
- Nonresponse: contact information from some units might not be accurate or available; others might not respond in time.
- Accuracy of response: some units might not think about the question on auditing in the same way as the researchers
- Usual random errors (typos, lost in mail, etc.)

Non response

- Unit non response – a unit/subject that is selected does not provide information
- Item non response – a unit/subject does not give information on a particular item/question
- Non-response bias occurs when respondents and non respondents are systematically different from each other (different on average) on an outcome of interest.
- Potential non-response bias can be addressed to some degree with survey weighting, statistical modeling, and imputation.

Non response and response rates

- Response rate: % of sample that responds – definition needs to be made clearly – see http://www.aapor.org
- Non response does not mean no information – longitudinal surveys, geographic information, etc.
- Incentives: incentives are sometimes used to encourage individuals to respond, especially to long, voluntary surveys. They hope to increase the response rate and thereby simplify analysis.
- Despite non response, often a survey can still give useful information about the population.

Common sampling designs in large-scale surveys

Simple random sampling is rare. Instead, surveys utilize stratification to control sampling and reduce standard errors. They use clustering to reduce costs. These affect sample selection and need to be considered in estimation.

Stratification

- Divide the population into subgroups. Take a sample separately within each and every group.
- Examples
  - Adults in households: stratify by state so that each state is represented.
  - Current Population Survey (CPS): 50 states plus DC, LA, and NY City
  - Census of Government hypothetical example: 2 government units per every state plus DC.
- Advantages: stratification can be used to increase precision and also force inclusion of specific groups

Weighting; Oversampling by strata

- Some groups are oversampled so that they get adequate representation for estimation within the group by itself
- Surveys usually have multiple outcomes of interest, so design is a compromise
- Estimates from samples within strata are combined as weighted averages to produce estimates for the population overall

Example: survey Congressional staff

- http://en.wikipedia.org/wiki/Congressional_staff
- 435 Representatives had on average 14 personal staff members (6090 total)
- 100 Senators had on average 34 personal staff members (3400 total)
- Stratify by chamber; sample 300 from each (SRS).
- Suppose average years in position is 3.8 in House and 4.6 in Senate. What is the estimated overall average?

$$\frac{6090}{6090+3400}\,3.8 + \frac{3400}{6090+3400}\,4.6 = 4.09$$

Example: survey Congressional staff

- 300 from each of the House and Senate is equal allocation of the sample to the two strata.
- Proportional allocation would give more to the House (385) and less to the Senate (215).
- Optimal allocation for the average overall likely would increase the number in the Senate assuming the standard deviation (variability) of number of years in position is higher in the Senate as well
- Equal allocation might be a decent compromise

Cluster sampling

- Picking households randomly over the U.S. or across a state could require a lot of travel time by interviewers.
- Instead, one can choose a random sample of groups of houses and interview households in those neighborhoods. This still is random and probability sampling, but it might not be as efficient statistically as simple random sampling.
- One can sample clusters within clusters: pick 100 counties, pick sub-county units within each selected county, pick households within sub-county units.

Sampling with probability proportional to size (PPS)

- Not all units contribute equally to a total, or average. They vary in size.
- One should be sure to include the largest units with higher probability. If units are very large, then they can be included with certainty.
- PPS sampling can be very efficient (small standard error). Note: you need to choose a measure of size.
- Largest population by county: Los Angeles County, California 9,818,605; Cook County, Illinois 5,194,675; Harris County, Texas 4,092,459; Maricopa County, Arizona 3,817,117; San Diego County, California 3,095,313.
- Largest land area by county: Five counties in Alaska

Combining sample design elements

- Usually a large-scale sample survey will include multiple elements to try to balance cost and efficiency
- The combination of effects are incorporated into survey weights so that the final weighted estimates are representative of the population.
- Estimation of standard errors becomes more complex and various statistical procedures have been developed to produce accurate estimates of standard errors in these complex designs.

Polls: factor affecting accuracy

Sample size affects sampling variability (bigger is better), but non sampling issues can be equally if not more important.

Example: Election political poll

- Target population: What is the population of interest? How is the target population identified?
  - Registered voters
  - Voters in last election
  - Those likely to vote in the next election
- Sample size: MOE for n=400 is 0.05; MOE for n=1000 is 0.03.
- Questions: What questions are asked? Wording and order of questions could make a difference.

Election political poll

- Frame:
  - RDD or white pages phone listing?
  - Are calls all made in the evening, or are they spread over days/times?
- Selection bias: Does each phone number get 1 call or many repeat calls? If only 1 call, then it is like a convenience sample.
- Response quality: Who is interviewing? How are they trained?
- Non response: What is the response rate? Which rate is being computed?

Election political poll

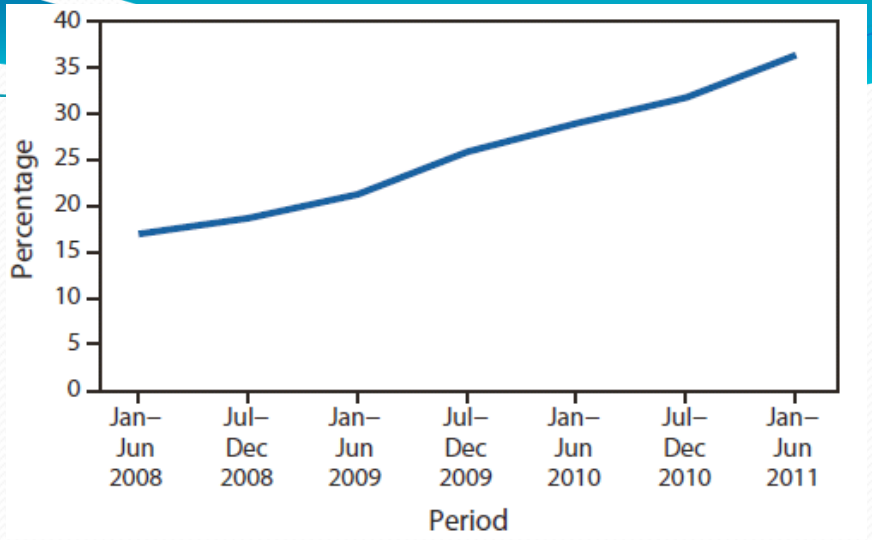
- Some polls use survey weighting to adjust for potential non response bias. How is survey weighting done? To what targets does the poll results adjust? How current and accurate is the population target information?
- If two or more polls disagree, then is there an explanation?
- Many of the questions asked about the accuracy of large-scale federal surveys can be asked about political polls.

Technology and Society

Telephones versus Paper

- In person interviewing is expensive, but it can be of high quality even for complicated survey.
- Telephone data collection replaced much in-person surveys as the percent of household with telephones rose well above 90%.
- Telephone is not cheap: you are paying someone to conduct the interview. A bigger problem is that response rates have dropped. This increases cost and hurts data quality.
- Mail survey data collection is less expensive, but quality can be lower (respondent can skip questions, get confused) and response rates can be low.

Cell phones



- CDC BRFSS: percent of cell-only household has risen dramatically. Not equally among all segments.
- Land-line telephone surveys are subject to potential coverage bias.
- Dual-frame surveys (land lines, cell phones) are being investigated
- Also there is a return to address-based surveys but now with multi-mode contact/response options

Internet

- Very useful as data collection mode
- Less useful as a frame for general population surveys; some segments of society have less access and usage.
- Various possible impacts on data quality:
  - Anonymous (positive)
  - Automated skip patterns and screen-optimized questions (positive)
  - Fast and inexpensive (positive)
  - Can't ask for help (negative)
  - Who is entering the data anyway? (negative)
- Future is in data collection through host of mobile devices: tablet, phone, etc.

Internet Panels

- Some organizations have recruited large numbers of volunteers to Internet panels.
  - After an initial contact, people are asked to join and provide background data.
  - Once someone joins the person is recruited to surveys.
  - Surveys can be tailored to specific topics.
- How do panels compare to the population? What groups are less represented? There can be quite a difference.
- What is response rate? It can be quite low.
- Can statistical methods adjust for differences? To some extent.
- How does one evaluate quality? Compare to Census or large-scale surveys done another way.

GIS and ADMIN

- Geographic Information systems (GIS): It is increasingly possible to link survey responses to exact locations.
- Administrative records: It is increasingly possible to link survey responses to administrative data.
- Both these developments can increase the value of information, provide checks on data quality, and potentially reduce cost of data collection.

Summary

Surveys and censuses are numerous and provide important information for policy research, evaluation, planning, and monitoring.

Survey sampling theory and methodology provide guidance for high-quality research that has been, is, and will be critical for most areas of policy, the economy, and society.

Statistical critical thinking can be very helpful to staffers as they confront data, studies, reports, and other information produced using sample surveys.

References

- ASA's What is a Survey: http://www.whatisasurvey.info/
- Survey Research Methods Section of ASA: http://www.amstat.org/sections/srms/
- AAPOR: http://www.aapor.org/
- Groves et al *Survey Methodology*, paperback book, Wiley, 2nd edition 2009
- Lohr, S. *Sampling: Design and Analyais*, hardback, Cengage, 2nd edition, 2009; more technical
- FEDSTATS: http://www.fedstats.gov
- CDC BRFSS – cell phones: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6122a3.htm

Contact

- Michael D. Larsen, PhD, GWU, mlarsen@bsc.gwu.edu
- Steve Pierson, ASA, spierson@amstat.org