

# Designing surveys

## Objective

- (1) Learn the technical jargon used in statistical surveys.
- (2) Understand the key issues that arise in survey design.

## Concept briefs:

There are many important terms here whose technical meanings differ from what you might guess through colloquial usage. Some of these terms are used extensively in later Chapters. So it is really worthwhile getting comfortable with them here.

- \* Sample vs. population - what is the difference?
- \* Statistic vs. parameter - what is the difference?
- \* Sampling strategies = (1) Simple Random Sampling (SRS); (2) Stratified random sampling; (3) Cluster sampling; (4) Multistage sampling; (5) Convenience sampling.
- \* Key bias problems = (1) Voluntary response bias; (2) Undercoverage bias; (3) Non-response bias; (4) Language or wording bias.

## Basic ideas in survey design

### Objective

You want to learn something about a target group - this is called the 'population.'

E.g., Who did EC students vote for in the last presidential elections?

### Key ideas

(1) You can get a good estimate by surveying a relatively small sample of your population.

(2) The main issue is to make sure your sample is "good."

(3) A good sample requires "randomizing" and picking an adequate sample size.

## How to design a survey: Basic recipe

(1) State the overall objective.

E.g.: We want to study the political leanings of EC students.

(2) Formulate specific objectives, or questions of interest.

E.g.: (a) Determine proportion of liberals, moderates & conservatives.

(b) Determine whether the proportions differ by sex.

(c) Determine whether the proportions differ by major.

(3) Identify the population & parameters.

E.g.: Population = EC students.

Parameters: (a) True % of L, M, C.

(b) True % of women L, M, C.

(c) True % of L, M, C, by major field.

(4) Sampling: Identify sample size, sampling method. Also analyze potential bias sources.

E.g.: Size = 100 women + 100 men;

Sampling method: stratified random. Select randomly from EC phonebook; send questionnaire by email & followup.

Potential bias: non-response; unlisted (?); other?

(5) Design questionnaire. Must typically have at least 2-3 questions for each major issue of interest. Unbiased wording is important. Most questions best to pose as multiple choice.

- 19. Mayoral race** Hoping to learn what issues may resonate with voters in the coming election, the campaign director for a mayoral candidate selects one block from each of the city's election districts. Staff members go there and interview all the adult residents they can find.
- 20. Soil samples** The Environmental Protection Agency took soil samples at 16 locations near a former industrial waste dump and checked each for evidence of toxic chemicals. They found no elevated levels of any harmful substances.
- 21. Roadblock** State police set up a roadblock to estimate the percentage of cars with up-to-date registration, insurance, and safety inspection stickers. They usually find problems with about 10% of the cars they stop.
- 22. Snack foods** A company packaging snack foods maintains quality control by randomly selecting 10 cases from each day's production and weighing the bags. Then they open one bag from each case and inspect the contents.
- 23. Milk samples** Dairy inspectors visit farms unannounced and take samples of the milk to test for contamination. If the milk is found to contain dirt, antibiotics, or other foreign matter, the milk will be destroyed and the farm reinspected until purity is restored.
- 24. Mistaken poll** A local TV station conducted a "Pulse-Poll" about the upcoming mayoral election. Evening news viewers were invited to text in their votes, with the results to be announced on the late-night news. Based on the texts, the station predicted that Amabo would win the election with 52% of the vote. They were wrong: Amabo lost, getting only 46% of the vote. Do you think the station's faulty prediction is more likely to be a result of bias or sampling error? Explain.
- 25. Another mistaken poll** Prior to the mayoral election discussed in Exercise 24, the newspaper also conducted a poll. The paper surveyed a random sample of registered voters stratified by political party, age, sex, and area of residence. This poll predicted that Amabo would win the election with 52% of the vote. The newspaper was wrong: Amabo lost, getting only 46% of the vote. Do you think the newspaper's faulty prediction is more likely to be a result of bias or sampling error? Explain.
- 26. Parent opinion, part 1** In a large city school system with 20 elementary schools, the school board is considering the adoption of a new policy that would require elementary students to pass a test in order to be promoted to the next grade. The PTA wants to find out whether parents agree with this plan. Listed below are some of the ideas proposed for gathering data. For each, indicate what kind of sampling strategy is involved and what (if any) biases might result.
- Put a big ad in the newspaper asking people to log their opinions on the PTA website.
  - Randomly select one of the elementary schools and contact every parent by phone.
  - Send a survey home with every student, and ask parents to fill it out and return it the next day.
  - Randomly select 20 parents from each elementary school. Send them a survey, and follow up with a phone call if they do not return the survey within a week.
- 27. Parent opinion, part 2** Let's revisit the school system described in Exercise 26. Four new sampling strategies have been proposed to help the PTA determine whether parents favor requiring elementary students to pass a test in order to be promoted to the next grade. For each, indicate what kind of sampling strategy is involved and what (if any) biases might result.
- Run a poll on the local TV news, asking people to text in whether they favor or oppose the plan.
  - Hold a PTA meeting at each of the 20 elementary schools, and tally the opinions expressed by those who attend the meetings.
  - Randomly select one class at each elementary school and contact each of those parents.
  - Go through the district's enrollment records, selecting every 40th parent. PTA volunteers will go to those homes to interview the people chosen.
- 28. Churches** For your political science class, you'd like to take a survey from a sample of all the Catholic Church members in your city. A list of churches shows 17 Catholic churches within the city limits. Rather than try to obtain a list of all members of all these churches, you decide to pick 3 churches at random. For those churches, you'll ask to get a list of all current members and contact 100 members at random.
- What kind of design have you used?
  - What could go wrong with your design?
- 29. Playground** Some people have been complaining that the children's playground at a municipal park is too small and is in need of repair. Managers of the park decide to survey city residents to see if they believe the playground should be rebuilt. They hand out questionnaires to parents who bring children to the park. Describe possible biases in this sample.
- 30. Roller coasters** An amusement park has opened a new roller coaster. It is so popular that people are waiting for up to 3 hours for a 2-minute ride. Concerned about how patrons (who paid a large amount to enter the park and ride on the rides) feel about this, they survey every 10th person on the line for the roller coaster, starting from a randomly selected individual.
- What kind of sample is this?
  - What is the sampling frame?
  - Is it likely to be representative?
  - What members of the population of interest are omitted?

1. **Playground, act two** The survey described in Exercise 29 asked: *Many people believe this playground is too small and in need of repair. Do you think the playground should be repaired and expanded even if that means raising the entrance fee to the park?*
- Describe two ways this question may lead to response bias.
2. **Wording the survey** Two members of the PTA committee in Exercises 26 and 27 have proposed different questions to ask in seeking parents' opinions.
- Question 1:** *Should elementary school-age children have to pass high-stakes tests in order to remain with their classmates?*
- Question 2:** *Should schools and students be held accountable for meeting yearly learning goals by testing students before they advance to the next grade?*
- Do you think responses to these two questions might differ? How? What kind of bias is this?
  - Propose a question with more neutral wording that might better assess parental opinion.
3. **Banning ephedra** An online poll on a website asked:
- A nationwide ban of the diet supplement ephedra went into effect recently. The herbal stimulant has been linked to 155 deaths and many more heart attacks and strokes. Ephedra manufacturer NVE Pharmaceuticals, claiming that the FDA lacked proof that ephedra is dangerous if used as directed, was denied a temporary restraining order on the ban by a federal judge. Do you think that ephedra should continue to be banned nationwide?*
- 65% of 17,303 respondents said "yes." Comment on each of the following statements about this poll:
- With a sample size that large, we can be pretty certain we know the true proportion of Americans who think ephedra should be banned.
  - The wording of the question is clearly very biased.
  - The sampling frame is all Internet users.
  - Results of this voluntary response survey can't be reliably generalized to any population of interest.
34. **Survey questions** Examine each of the following questions for possible bias. If you think the question is biased, indicate how and propose a better question.
- Should companies that pollute the environment be compelled to pay the costs of cleanup?
  - Given that 18-year-olds are old enough to vote and to serve in the military, is it fair to set the drinking age at 21?
35. **More survey questions** Examine each of the following questions for possible bias. If you think the question is biased, indicate how and propose a better question.
- Do you think high school students should be required to wear uniforms?
  - Given humanity's great tradition of exploration, do you favor continued funding for space flights?
36. **Phone surveys** Any time we conduct a survey, we must take care to avoid undercoverage. Suppose we plan to select 500 names from the city phone book, call their homes between noon and 4 P.M., and interview whoever answers, anticipating contacts with at least 200 people.
- Why is it difficult to use a simple random sample here?
  - Describe a more convenient, but still random, sampling strategy.
  - What kinds of households are likely to be included in the eventual sample of opinion? Excluded?
  - Suppose, instead, that we continue calling each number, perhaps in the morning or evening, until an adult is contacted and interviewed. How does this improve the sampling design?
  - Random-digit dialing machines can generate the phone calls for us. How would this improve our design? Is anyone still excluded?
37. **Cell phone survey** What about drawing a random sample only from cell phone exchanges? Discuss the advantages and disadvantages of such a sampling method compared with surveying randomly generated telephone numbers from non-cell phone exchanges. Do you think these advantages and disadvantages have changed over time? How do you expect they'll change in the future?
38. **Arm length** How long is your arm compared with your hand size? Put your right thumb at your left shoulder bone, stretch your hand open wide, and extend your hand down your arm. Put your thumb at the place where your little finger is, and extend down the arm again. Repeat this a third time. Now your little finger will probably have reached the back of your left hand. If your arm is less than four hand widths, turn your hand sideways and count finger widths until you reach the end of your middle finger.
- How many hand and finger widths is your arm?
  - Suppose you repeat your measurement 10 times and average your results. What parameter would this average estimate? What is the population?
  - Suppose you now collect arm lengths measured in this way from 9 friends and average these 10 measurements. What is the population now? What parameter would this average estimate?
  - Do you think these 10 arm lengths are likely to be representative of the population of arm lengths in your community? In the country? Why or why not?
39. **Fuel economy** Occasionally, when I fill my car with gas, I figure out how many miles per gallon my car got. I wrote down those results after six fill-ups in the past few months. Overall, it appears my car gets 28.8 miles per gallon.
- What statistic have I calculated?
  - What is the parameter I'm trying to estimate?
  - How might my results be biased?
  - When the Environmental Protection Agency (EPA) checks a car like mine to predict its fuel economy, what parameter is it trying to estimate?

### **Example: Ex. 19, p. 288**

#### **Solution:**

Population - City voters.

Population parameters - Not clearly indicated. Possibly % of voters who support (or oppose) various specific issues.

Sampling frame - All residents in the city.

Sample - All residents they can find in 1 block from each election district of the city.

Sampling method - Multistage cluster (and convenience?) sample. No randomization used. The staff interviews every resident they can find.

Left out - Anyone not home at the time when the surveyors visit.

Bias or problems - Sampling within clusters is not random, so it may bias results. Also, parameters of interest have not been thought out, or clarified, which is a potential source of problems in how they conduct survey.

### **Exercise 21, p. 288**

#### **Solution:**

Population - Cars within a city or region (details not clear).

Population parameters - Number (or proportion) that don't have up-to-date registration, insurance or safety inspection.

Sampling frame - All cars on that particular road.

Sample - All cars that pass through that particular roadblock.

Sampling method - Kind of cluster, by area of city. Perhaps there is randomization in selection of roadblock location & time (details not clear).

Left out - Cars not driving on that particular road, or those driving on the road at some other time.

Bias or problems - Randomization of location & time of day is important. If they haven't done that the results may not be representative.

**Exercise 34-35, p. 289**

**Solution:**

(34.a) Words that introduce bias: “pollute” and “compelled.”

(34.b) Response biased toward “No” because of comparisons with military & voting eligibility.

(35.a) No obvious bias.

(35.b) Again, use of extraneous words & comparisons before asking the question.

**Exercise 39, p. 289**

**Solution:**

(a) The average (mean) gas mileage of my car for recent 6 fill-ups.

(b) Mean gas mileage of my car.

(c) The recent 6 fill-ups may not be representative (or, typical).

(d) The mean gas mileage of all cars same as my car’s make & model.