Biased Samples – Solutions
STAT-UB.0103 – Statistics for Business Control and Regression Models

Populations and Bias

Each of the following scenarios involves collecting data to learn about a population. State (a) what population is involved, and (b) why the sample is biased. To demonstrate that a sample is biased, you must argue that certain members of the population are more or less likely to be sampled than others. Note: there will usually be many valid answers for parts (a) and (b), but your answer to part (b) will depend on how you define the population in part (a).

1. You need a survey on household spending patterns. You take a random sample from the customer list of the local brokerage firm.

   Solution: *Population: the spending patterns of all households.* People who have brokerage accounts are certainly more wealthy than average, so richer people are more likely to be sampled.

2. You want to learn about New York City residents' sentiments (positive or negative) towards their new mayor, Bill de Blasio. You search for "de Blasio" on Twitter and read the first 100 relevant search results.

   Solution: *Population: the sentiments towards de Blasio for all New York City residents.* People who use twitter are more likely to be sampled than people who don't. People who have tweeted about de Blasio are more likely to be sampled. Furthermore, even among all people who use twitter, and have tweeted about de Blasio, those with recent tweets are more likely to be sampled.

3. You need to know the opinions of Stern students with regard to some curriculum matters. You ask some of the people in your class.

   Solution: *Population: the opinions of all Stern students.* Sophomores, Juniors, and Seniors are less likely to be sampled, as are people who prefer morning classes.

4. You want to learn about the quality of the food at a local restaurant. You read the reviews for the restaurant on Yelp.com.

   Solution: *Population: opinions of all people who have eaten at the restaurant.* The people who write reviews on Yelp.com are more likely to be sampled than people who do not use Yelp, or do not have Yelp accounts. (Yelp reviewers tend to have extreme opinions towards food, or tend to be overly critical.)

5. You want to estimate the rate of growth of stocks over the last fifty years. You take a random sample of the stocks listed today on either the New York Stock Exchange or the Nasdaq. Some of these stocks did not exist fifty years ago; you set these aside. For the other stocks, you identify their prices fifty years ago, and you use this to compute the growth rate.

> **Solution:** *Population: the rates of growth of all stocks over the last fifty years.* Companies that were listed fifty years ago but did not survive are not available to appear in your sample. This is an example of survival bias. You will seriously overestimate the growth rate!

6. You want to know information about consumer preferences on a number of household products, including soap, laundry detergents, dishwashing detergents, furniture polish, and cleanser. You devise a questionnaire item with 50 questions; this takes ten to fifteen minutes to administer over the phone. You randomly select phone numbers, and you get the responses of those who are home and willing to help you.

> **Solution:** *Population: the opinions of all consumers.* You have here a bias in favor of people who are at home and answer their phone. Such people may have non-typical opinions about consumer products. (Even if you were dealing with a non-home related topic, such as recent movies, you would still have a biased sample.) Finally, you are only getting the opinions of people who are willing to waste ten to fifteen minutes of their time talking to you! Why do we care about the opinions of such people?

7. You want to know whether a certain teaching method improves the reading abilities of fourth-grade students. You examine all the articles on this subject published in five major education journals in the last ten years.

> **Solution:** *Population: the reading ability improvements of all fourth-grade students exposed to the teaching method.* Journals have a prejudice toward publishing articles which show strong statistical relationships. Submitted articles which show that the method fails are not likely to be published. This is called *publication bias*.

8. You want to learn about lifestyle habits which lead to kidney cancer. You take a random sample of patients from the list of an oncology practice, and you interview these people with regard to issues like diet, cigarette smoking, chemical exposure, and so on.

> **Solution:** *Population: The lifestyle habits of all people who get kidney cancer.* This is clear survivor bias. You are more likely to sample cancer survivors and people who live with kidney cancer for longer periods of time.

9. You want to get information about some mutual funds, so you research every fund which was advertised in the last four issues of a financial newsletter.

Solution: *Population: information about all mutual funds.* This is a variant on publication bias. You will only get to see ads from funds that have done very well in the recent past.

10. You want to learn opinions among parents in your school district regarding adult literacy education. You send out a letter inviting all parents inviting them to attend an information session.

Solution: *Population: opinions of all parents in your school district.* This is selection bias. Obviously the illiterate will not be reading this letter about the meeting!