We will often criticize a set of data as being biased. Perhaps we’ll say that it was collected in a biased way.

Here we mean either or both of these:

* Not every item from the target population is equally likely to come to our attention.
* The items that show up in our database are not obtained independently. By this we mean that the occurrence of a particular item makes the occurrence of another particular item more (or less) likely.

We often say that biased samples are non-representative, meaning that the sample does not speak correctly about the population in some sense. This would be a bit misleading, as “bias” and “non-representativeness” are not exactly the same thing.

* Completely proper data samples can end up being non-representative just by bad luck.
* Biased samples can still be representative.

Example: You want to compare Diet Coke and Diet Pepsi. Instead of properly selecting a sample, you just ask the first 20 people you encounter. This is a biased sample, but it could easily be representative in terms of the question about the two drinks.

You should avoid biased sampling processes. These usually lead to situations of non-representativeness. The bigger problem is that biased methods do not enable you to make good generalizations.

There are quite a few styles in which we end up with biased data.

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

This is easy. People who have brokerage accounts are certainly more wealthy than average, so your sample is biased in favor of richer people.

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

Your sample might not be representative. After all, these people are taking the same course you are! Also, there may be serious differences between daytime and evening students.

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.

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.
You want to estimate the rate of growth of stocks over the last fifty years. You take a random sample of the stocks listed 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.

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!

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.

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?

There are similar issues related to volunteer postings on web sites for movie reviews, book reviews, and so on.

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.

This is clear survivor bias.

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. 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.

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

This is **selection** bias. Obviously the illiterate will not be reading this letter about the meeting!