

## Populations, samples, surveys and statistical inference

The essence of statistics is **inference** — taking available information, and inferring properties of the objects that we did not see. This is fundamentally connected to the ideas of *populations* and *samples*.

**Population:** the population is the entire group of entities that we are interested in examining.

**Sample:** the sample is a subset of the population that is the actual subject of examination. We can learn almost anything about the sample, but wish to learn corresponding things about the population.

Why do sampling at all?

- (1) cost
- (2) it can be impossible to “get” everyone in the population
- (3) it can be more accurate (!)

Many samples (particularly of people) are obtained by taking a survey, but not all samples are best obtained this way. Before starting, it is important to consider if the required information can even be collected by a survey; sometimes running an experiment is better, and sometimes it is only possible to measure quantities of interest indirectly. If a survey is decided upon, the first step is to lay out the objectives of the investigation — what do you hope to find, and how will you measure it? The objectives should be as specific, clear-cut and unambiguous as possible, subject to the inevitable tradeoffs that exist. Will the survey be by mail, telephone or in-person?

The key to a good sample is that it be typical of the population from which it is drawn. When the information from a sample is not typical of that in the population in a systematic way, we say that **bias** has occurred. It is only common sense to require that concepts in any survey be clearly defined and questions unambiguously phrased, in order to minimize misleading conclusions. There are many different kinds of bias; here are a few of the common causes of bias (often from surveys):

- (1) *selection bias* — subgroups of the population are systematically under or overrepresented in the sample
- (2) *self-selection bias* — sample membership is determined by the respondents themselves, such as in telephone dial-in convenience samples
- (3) *nonresponse bias* — the subjects that respond are different from those that do not (this is similar to selection bias, although a distinction is that it is possible that the correct respondents have been sampled, but refuse to respond)

- (4) *response bias* — the responses of subjects do not actually reflect their true opinions, because of outside pressure, ambiguity of wording of a question, the desire to please the interviewer

A pretest, or pilot study, is the only way to find out if the questionnaire and field procedures work correctly. Because it is rarely possible to foresee all the potential misunderstandings or biasing effects of different questions and procedures, it is vital for a well-designed survey operation to include provision for a pilot study (or series of pilot studies). These are used to test the feasibility of techniques and to perfect the questionnaire concepts and wording.

The underlying principle that must be followed if we are to have any hope of making inferences from a sample to a population is that the sample be representative of that population. A key way of achieving this is through the use of “randomization,” where (in a clinical trials situation, for example) treatments are assigned at random to patients, and members of the population are chosen to be in the sample randomly. Randomization puts systematic effects and potential biases into the “noise,” where they can’t be misinterpreted as “signal.” Random assignment of treatments also provides justification of statistical methods free of restrictive assumptions. Indeed, assigning treatments randomly is much more important than splitting them evenly over the sample (achieving balance). For example, in a study of a rare disease, the number of “cases” (people with the disease) is likely to be much smaller than the number of “controls” (people without the disease), but this is not problematic.

There is one kind of error that is not a problem — **sampling error**. This is the error that comes from the sample not being the same as the population. In a well-designed experiment or survey, these errors can be estimated well.

“By a small sample we may judge the whole piece.”

— Miguel de Cervantes

“‘I was counting the waves,’ replied Amory gravely, ‘I’m going in for statistics.’”

— F. Scott Fitzgerald, *This Side of Paradise*

“Get the facts first, and then you can distort them as much as you please.”

— Mark Twain

“To guess is cheap, to guess wrongly is expensive.”

— Chinese proverb

“That which is separated from a collection derives from the majority.”

— The Talmud (*Mishne Torah: Tefillin*)

“‘Multiple births are more frequent in larger families,’ declares a statistician. It’s mighty hard to fool these statisticians.”

— Unknown

“Collecting data is much like collecting garbage. You must know in advance what you are going to do with the stuff before you start collecting it.”

— Mark Twain

“Get it right or let it alone.

The conclusion you jump to may be your own.”

— James Thurber, *Further Fables for Our Time*

“Not everything that can be counted counts, and not everything that counts can be counted.”

— Albert Einstein

“The king lost his way in a jungle and was required to spend the night in a tree. The next day he told some fellow traveler that the total number of leaves on the tree were ‘so many’ (an actual number was stated). On being challenged as to whether he counted all the leaves he replied, ‘No, but I counted the leaves on a few branches of the tree and I know the science of die throwing.’”

— From the ancient Indian epic Mahabharat

(*Nala-Damayanti Akhyān*)