# Collecting data
## Math 218, Mathematical Statistics

D Joyce, Spring 2016

Begin a discussion of collecting data. Topics include types of statistical studies, control groups in a comparative study, sample surveys, prospective and retrospective studies, basic sampling designs, simple and other kinds of random sampling. (Chapter 3)

**Simple random sampling.** Simple random sampling gives values of a probabilistic random sample, that is, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ where the $X_i$'s are independent random variables which all have the same distribution. Theoretically, this situation is easy to treat, and that's what we'll do. But practically, such a random sample is hard to create. Usually the values we get from a statistical study are not independent and don't have the same distribution. Nonetheless, you have to work with the data you have.

**Types of statistical studies.** One kind of statistical study depends on data that's already recorded, called historical data. For instance, if you're interested in determining whether there is a global warming trend, one thing you can do is look at the historical record of temperatures. A study of historical data usually involves some kind of "time-series" analysis. We'll survey that in the next chapter.

A very common kind of statistical study is one which compares two groups or methods. Such comparative studies analyze one or more attributes of these groups or methods.

Studies may be observational or experimental. If a study is experimental, then the groups or methods are controlled. Often, control is not possible, but observations can still be made.

Whatever the kind of study, various characteristics are measured. Those characteristics are called *variables*, and their measured values called the *data*. Some of these variables are called *predictor* or *explanatory variables*. They're something like the independent variables in ordinary algebra. For instance, variables that can be set by an experimenter are predictor variables. But even when none of the variables can be set, one or more of the variables can be selected to be the predictor variables. The remaining variables are the *outcome* or *response variables*. The goal is often to see how, or if, the predictor variables determine, wholly or in part, the outcome variables.

In many situations, there may be other variables—that may or may not be measured—that have more effect on the predictor variables. Such variables are sometimes called *confounding variables*, and, in the case that they aren't recognized until after the fact, *lurking variables*.

In any case, a statistical study does not show cause and effect.

**Control groups in a comparative study.** Control groups are often used to reduce confounding variables. Subjects are divided into two groups—those given a treatment under study, and those that aren't. Placebos and single blind tests hide what's being tested from the subjects of the test so they don't know which group they're in. Double blind tests also hide that information from those performing the tests, too.

**Sample surveys.** In a survey, there is some entire population of items—humans, animals, objects, or whatever. A numerical characteristic of a population is called a *parameter*. When the entire population can be measured, that's called a *census*. Frequently, only part of the population can be measured, and that portion of the population that is actually measured is called a *sample*. The numerical characteristic for that sample is called a *statis-*

*tic.*

Of course, the purpose of the statistic is to answer questions about the parameter. In order to improve the power of the statistic, the sample should be *representative*, and one way to do that is to make it a *random* sample. (See below.)

**Prospective and retrospective studies.** Many studies, of whatever kind, are made over time. Those that start at the present and go forward are called *prospective studies*, those that use past information are called *retrospective studies.*

Studies that follow a group of individuals over time are called *cohort studies.*

**Basic sampling designs.** A crucial part of the design of an survey is the way that the sample is selected from the population. As mentioned above, it should be representative, that is, it should not differ in systematic or important ways from the population, except, of course, it is much smaller. Poorly chosen samples are may be biased and lead to invalid conclusions about the population.

**Simple random sampling.** If all the individuals in a population have the same probability in being selected for the sample, then that sample is called a *simple random sample.* If the entire population is $N$, and the sample size is $n$, then there are $\binom{N}{n}$ possible samples, and each individual in the population has a probability of $n/N$ of being part of a sample. The value $n/N$ is called the *sampling rate* or the *sampling fraction.*

This is actually sampling without replacement so the hypergeometric distribution is appropriate, but the sampling rate is usually so small that sampling without replacement is almost the same as sampling with replacement, so the hypergeometric distribution can be replaced a binomial distribution making the theory and practice easier to do.

The methods we develop in this course assume that we have a simple random sample.

**Other kinds of random sampling.** Sampling is expensive and many ways have been developed to reduce the size of the sample yet still get good information.

Quota sampling is sometimes used to guarantee that a certain aspect of the population is fairly represented, but other aspects, which may be more important, can be overlooked.

If we combine quota sampling and simple random sampling together, we get someting called *stratified random sampling.* The entire population is divided into known subpopulations, then a random sample is selected from each subpopulation. With stratified random sampling, correlations between the stratified aspect and lurking variables can lead to invalid results.

A variant of stratified random sampling is *multistage cluster sampling.* When an entire population is partitioned into subpopulations, first simple random sampling is used to select a certain number of those subpopulations, then for each of the selected subpopulations simple random sampling is used to select individuals. There can be more than two stages, too.

A technique that is often used in place of strict simple random sampling is called 1*-in-k systematic sampling* in which every $k^{\text{th}}$ individual in a list of the population is chosen.

**Experimental studies.** In the terminology of experimental design, purposes of an experiment are to

- evaluate how a set of *factors* (that is, predictor variables) affect one or more response variables
- screen out unimportant factors
- select values for the controlable factors to maximize or minimize a response variable
- fit a model that can be used to make predictions or adjust controlable factors to determine particular values for one or more response variables.

The first purpose, of course, is the primary purpose.

Those factors that can be controlled are also called *treatment factors*. Other factors, some of which may not even be known, are called *noise factors* or *nuisance factors*. The values of a factor are also called its *levels*. A *treatment* is a particluar combination of levels for all the treatment factors.

Treatments are applied to *experimental units* (i.e., subjects), and measuring the responses. A *treatment group* is a collection of experimental units who all get the same treatment, and the result of that treatment on the treatment group is called a *run*.

See example 3.8 which illustrates these concepts.

**Strategies to reduce experimental error variation.** There are various kinds of experimental errors. One is *systematic error* due to differences between experimental units; various units can be biased to give nonrepresentative responses. Another is *random error* which is the variability of responses that an individual unit will give. A third is *measurement error* which means the same response may be measured differently if a new set of measurements are made.

See example 3.9.

There are some standard strategies that are used to reduce systematic error.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`