Introduction to the Course and to Statistics

What is This Course About?

This course is meant to provide an introduction to statistical methods, with a concern of highlighting their scope and limitations.

What Are Our Main Goals?

We have a limited time frame available, and, though the technicalities of the statistical tools we will study are not very difficult (the actual calculations involved are mostly arithmetical, and the more difficult ones are handled by tables and software), the effort to get in the spirit of statistical analysis is non trivial, and will take up enough of our time. Hopefully, this will make us better **consumers** of statistics (and we are flooded with statistics every day, so that being an educated consumer is definitely worthwhile). It will also allow us to perform several common statistical analyses on data sets, which were hopefully collected using best practices. We will discuss what best practices are meant to produce, even if we will not be able to get into really operational details.

What Are We *Not* Going To Learn Here?

We cannot and do not expect to become a professional statistician after a quarter-long course. For this reason, not much emphasis will be put on the actual planning and executing of a statistical experiment. We will discuss some practical methods in general terms, but the further step to implement these methods effectively is too complex a subject to be covered in a few pages.

Again, best practices are meant to produce data sets satisfying a number of requirements. We will discuss what these requirements are, and sketch some of the methods that have been devised to satisfy them. However, bear in mind that "design of experiments" is an active area of research, and it is full of pitfalls that could be easily missed. Still, realizing what makes a "good" data set will help you in assessing whether a given list of numbers passes muster or is suspect.

What is Statistics?

The root of the name is “State”: its origins lie in the need for governments to gain an accurate picture of resources available to their country. This is quite a different problem from traditional censuses, which have been held “forever”. *Census* derives from a Latin word whose root means “money”: a census was a rough head count to determine how much tribute and tax rulers could hope to extract from their subjects. Precision was not a high priority in such cases.

As the modern nation-state emerged in Europe and then in North America, the goal turned to getting data in order to promote as much economic development as feasible, and that required both much more information as well as much better accuracy. The “accuracy” part is really tricky, as became clear very soon. The account of how this quest eventually led to our modern discipline, heavy on mathematics, that we call “Statistics” is very interesting, and you are urged to check it out from the many books that have been written on the subject.

Census

There have always been censuses, but the modern idea is very different, as sketched above. In principle, it is supposed to get a precise count of the population, as well as information on characteristics, resources, and so on. Thus, in theory, it provides a “perfect” snapshot of a country. In practice, since it is an expensive endeavor, it cannot be done continuously, but, additionally, the “perfect” is impossible to achieve. This impossibility calls for statistical methods to deal with it, but they are outside our scope in this course, as they are very specialized to this unique environment.

This is not true, of course, if the population of interest is small and well defined. If you are going to study your music collection, at a fixed date, even if you have a huge amount of music pieces, the "population" is still manageable with a computer program, and, additionally, it is well defined (you either have a piece of music in the collection now, or you don't). In this case, your observations will describe fully your object of interest.

Sampling

The more interesting application of statistics is, however, when you are dealing with a very large, if not potentially infinite, "population", and are only able to observe a relatively small set of elements. In this case, you are not really interested in the limited number of observations, per se: what you are looking for is a way to reach conclusions about the general population, from the limited data you have.

What then?

For a while, up to the 19th Century, the solution was thought to call for a search for an elusive "typical" element - town, citizen, you name it. This "typical" object would have the feature of providing a picture of what was "typical" of the country. It dawned very soon that this could not work, first of all because of the vagueness of the requirement, and, in a related fact, because of the practical impossibility of determining such a "typical" object in any objective way. Finally, in the early 20th Century, thanks to brilliant researchers, such as Fisher, Pearson, Neyman, and many others, a solid logic foundation was set for quantitative treatment of partial data, from a completely different point of view: probability.

Examples

Rather than set out abstractly, let us look at a few typical situations where statistics is used to study large populations from limited data:

→ "Polling": out of a large population (say, the population of. the United States), a few individuals are chosen and their answers to questions are recorded. Somehow, we would like to extrapolate conclusions about **the general population** out of these few individuals.

→ "Measuring": a physicist measures a physical quantity (e.g., the speed of light). She repeats the experiment several times, and the numbers she observes are all different. Somehow, out of these numbers, we would like to extract the "true" value of the speed of light. Note that, in principle, there is no limit to the number of times this experiment could be performed, and hence on the number of different numbers we may observe.

→ "Survival Analysis": here again, we measure something, but this time it is the time to failure of a product (e.g., how long does it take for a light bulb produced in our factory to burn), or the life spans of living beings. Since this is a destructive measurement, it cannot be performed on every item produced, but only on a few samples. Somehow, we want to deduce an estimate for the time to failure of **any** of the products. In this case, depending on the circumstances, we may want to apply our deductions to the production batch from which our sample was drawn, or even to the whole production line. In the latter case, again, the "population" is potentially unlimited in size.

→ "Correlation": we observe two quantities that we suspect influence each other (e.g., level of education and income), but only for a limited number of cases. Somehow, we want to deduce a general connection between these quantities.

Please, consider these example carefully: each one has you looking at some numbers, but, in principle, none of the conclusions that we would like to draw has any basis. To proceed, we have to assume something about the connection between what we observed, and what we would like to actually learn. This connection is not automatic: it is a model that we adopt, in the hope that it is a realistic model. This necessity is basic for understanding what the various techniques we have to "describe" the data mean, and what their validity is. The type of model that has proved to be the most useful for our needs is a probabilistic model. As we will see, assuming such a model is a reasonable description of the actual procedure we followed in collecting our data, we will be able to draw quantitative conclusions for our questions. However, if, by any chance, the data was collected in a way that does not match our model, any application of the techniques we will learn is completely meaningless, and any "conclusion" we might try to draw is baseless.

Descriptive Statistics

Any statistical activity begins by collecting data. Since this usually means collecting a fair amount of information, it is useful to be able to summarize this information into digestible chunks. Suppose, for example, that, following the work of the Scottish pioneers of the insurance industry, you collected a long list of life spans of many people from some data base. Dealing with hundreds, thousands, or even more, numbers is not something humans do well, so some way of organizing the numbers that allows us to grasp how they look is necessary. Over the years, a number of methods have been devised to this effect. Whatever method you choose, however, you should bear in mind that this is **a way to summarize the data you have**. It does not imply any further conclusion. Thus, a histogram of the lifetime data we mentioned will be easier to read than a long table of numbers, but its value is restricted to the data it summarizes. To go beyond the raw data and to draw more general conclusions requires appropriate tools, which will be our topic when we move to *Inferential Statistics*. Another fundamental point to remember is that in most cases any summary comes at the cost of *some loss of information*. What's more, there is no unique way to summarize with loss of information, so a critical eye is needed to avoid possible serious misunderstandings.

Exact Small Populations

Sometimes the data refer to a complete information set. As an example, consider an instructor who teaches a class of about 35 students every semester. Each time she might give an entrance test to assess the level of preparation of her students. Also, she might keep track of the results of subsequent tests, and of the final outcome of the class. Since all of this produces a significant body of numbers (let's assume grades are numerical in her class), and humans are very bad at extracting information out of a big list of numbers, she will use some of the summarizing techniques that we will discuss later. This is all good, and is something that could even be done "by hand", as in using a calculator instead of a computer to work out the numbers. While this would be an acceptable use of statistical tools, it is definitely not the most interesting or even useful. Fact is, in this case all the instructor is doing is learning about the specific 35 students in one specific class, but no general statement beyond her class is warranted.

Samples From Large Populations

Since, as we may suspect, there is a temptation to extrapolate from the students in our teacher's class to a broader population (as in "from my observations, I see that students come in less prepared than they used to", implying that what was observed on a few particular groups of 35 students is true for the remaining millions out there), one goal of this course is to make clear that such extrapolations are unwarranted without further assumptions (even if we all do it all the time). To deduce general conclusions from limited observations, we need a lot of safeguards and methodologies in place, and none of these would be applicable in our scenario. To apply limited observations to general statements, we need a solid model of how this would be reasonable, and we need the actual experiment to be reasonably conforming to the model.

Once this is in place, we can actually do it, and that's what we will be concentrating on in this course. Whenever the data refers to a limited sample, descriptive tools only tell us what the sample looks like, but do not imply anything about the broader population. To jump out of the limited sample, we need a *model*, that formalizes how, and how far, data about a sample can suggest information about the whole population. In other words, a model will provide a theoretical connection between the sample and the population. We will then need to make sure that the real practical method used to obtain the sample is reasonably well described by the theoretical connection described by the model. There are so many examples of faulty application of this approach that it is important to have this issue very clearly in mind, both when working on your own statistical experiment, and when looking at other people's work.

Inferential Statistics

The methodology that is generally used to draw general conclusions from limited observations goes by the name of *Inferential Statistics*. It is a collection of methods based on *probability models*, that allow us to assert facts about the general population, based on our limited observations, *with a given degree of plausibility*. No statistical statement asserts certainty. We can assign a quantitative measure of certainty to statistical statements, but this measure is never 100%, except in completely trivial cases. Nonetheless, if his degree is sufficiently high, it is a reasonable basis for further decisions, and it certainly beats decisions based on instinct or subjective evaluation of the facts.

Of course, the quantitative element *depends on the specific probability model we adopt*. While in many cases this choice is not difficult, that is not always so. A currently fashionable example is given by the risk models used by banks and financial institutions in the years leading to the market crash of 2007. These models used statistical data on market behavior and applied a common class of models, called *normal* (we will be mostly concerned with such models in our practice). However, it turned out that these models grossly underestimated the risk of large market fluctuations, with very dire consequences when these fluctuations actually occurred. A number of researchers have argued (already before the crash) that a more sensible toolbox would have been provided by so-called "fat tail" models (a simple, stylized, such model is quickly illustrated as one of the "special topics"). Ironically, the very same diagnosis was put forward in the wake of the disastrous failure of the *Long Term Capital Management* hedge fund in 1998 (the Federal Reserve had to intervene directly to prevent this failure to propagate and generate a general market disaster). Apparently, that experience was not enough to make people wary of "rare event" risks.

Outline Of Our Course

In this course we will cover many topics. The general scheme is as follows.

Summarizing Observations

For small, well-defined populations, collecting a complete data set is a feasible project. More commonly, we are going to deal with a limited set of data, compared to the whole scope of our investigation. In any case, it is almost always necessary to find ways to summarize the data, as we find it difficult to grasp a long list of numbers or other data. The key here is that all we can do is **summarize the data we have**. No attempt should be made, with these tools only, to draw conclusions beyond the specific group of observations we have.

Probability Models

Any hope of generalizing from limited observations relies on being able to apply a mathematical model as developed in Probability Theory. We will take a short tour of this theory, without any pretense of completeness or strong rigor, but with the goal of acquiring some important conceptual tools, as precisely as possible.

Inferential Statistics

The application of Probability models to partial data collections is called Inferential Statistics. We will look at a few of the main ways in which this can be done. This is a vast subject, and we can only cover a small part of it.

Estimation

This area starts with the choice of a *class* of probabilistic models, and uses the data, together with theory, to reduce the class to a limited number of choices, that seem most compatible with the actual observed data

Testing

Here, the data is used to test whether the model we think is appropriate is reasonably consistent with the observations. To make this procedure really useful, we will also look at how our chosen model performs compared to selected alternative ones.

Applications

After having learned about the theory, we can take a look at how it is applied in practice. It is a difficult area, if tackled right. We would want to keep the gap between the abstract requirements of our mathematical models, and the practical actions for collecting data to a minimum, but there are many possibilities for failure. This issue is made even worse when, for one reason or another, the data collection was flawed form the start. This should be familiar: we have all heard of statistical analyses that proved to be faulty, and, actually, not all of them were due to shoddy practices.