Statistics: introduction

S. Maset

Dipartimento di Matematica e Geoscienze, Università di Trieste

DDM PEM 2018-2019

Outline

Introduction

- In order to learn about something happening in the real world, we need to collect data.

  Example: we want to know whether or not the age differences of the children, when they begin the primary school, have some influence in their scholastic career.

  People do not like collecting and analyzing data, they prefer to remind one or two cases (anecdotes) and then they begin to speak and to argue.

  The conclusion of someone could be the following: the differences of ages are important because, at the age of six years old, some months of difference provide a different maturity.

  The argument seems to work: older children understand better what is taught in the first school year and then, since what is taught at the first year is fundamental, they will have a better scholastic career.

Indeed, it is confirmed by tests at the end of the first school year that the older children have understood better what has been taught. But, can the younger children fill the gap in the successive years?

What do the data say? From the US census, we have

Table 1.1 Total Years in School Related to Starting Age

Year	Younger half of children		Older half of children	
	Average age on starting school	Average number of years completed	Average age on starting school	Average number of years completed
1946	6.38	13.84	6.62	13.67
1947	6.34	13.80	6.59	13.86
1948	6.31	13.78	6.56	13.79
1949	6.29	13.77	6.54	13.78
1950	6.24	13.68	6.53	13.68
1951	6.18	13.63	6.45	13.65
1952	6.08	13.49	6.37	13.53

Source: J. Angrist and A. Krueger, "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, vol. 87, no. 18, 1992, pp. 328–336.

Conclusion: the age differences are not important, at least in US and for the number of years of school completed.

Exercise. Are the age difference also important for sports? Do an internet research about this question by considering football.

- In the previous example, we have learned something from data.

  Statistics is the art of learning from data.

  Statistics is concerned with:

  - the collection of data;
  - the description of the collected data;
  - drawing conclusions from the collected data.

  The word "Statistics" comes from "State", because the states historically were the first to collect data on their populations.

The collection of data

- In some situations, we have at our disposal a given set of data. For example, in the previous situation of age differences in beginning the primary school, we have data from the US census.

  In other situations, the data are not yet available. In this case, we need to design an experiment to generate the data.

  Example: suppose that a new cholesterol-lowering drug has been developed and its efficacy needs to be determined. To this aim:

  - *some volunteers are recruited and their cholesterol levels are measured;*
  - *we give to the volunteers the drug for some period of time;*
  - *then we measure again their levels of cholesterol.*

  If the cholesterol levels are significantly reduced, then we conclude that the drug has efficacy.

However, the experiment is not well designed. The reduction of cholesterol levels can be caused by other reasons: for example

- *some placebo effect that induces a more healthy lifestyle;*
- *warm weather with a consequent increase of the physical activity caused by the stay more time outdoors.*

A better experiment should try to neutralize all possible causes of the change of cholesterol level different from the drug.

How to do this?

Divide the volunteers into two groups:

- *one group receives the drug;*
- *the other group, called the* ***control group****, receives a placebo that has the same look and the same taste as the drug, but has no physiological effect.*

The following three advices should be considered.

- *The volunteers should not know whether they are receiving the true drug or the placebo. We want to avoid placebo effects.*
- *The medical people overseeing the experiment should not be involved in the division of the volunteers into the two groups and should not know who are receiving the drug or the placebo. We want to avoid frauds: suppose that the doctors that created the new drug know that the drug has no effects on smokers; the doctor could select only non-smokers in the group receiving the drug or only smokers in the control group.*
- *Finally, in order to avoid any biased behavior, the division of the volunteers has to be done at random.*

If these three advices are implemented, the experiment is called a ***double-blind randomized trial****.*

The description of the data

- The part of Statistics concerned with the description and summarization of collected data is called **Descriptive Statistics**.

  In the example of cholesterol-lowering drug:

  - description: the cholesterol levels of each volunteer, before and after the experiment, should be presented, along with the information whether the volunteer has received the drug or the placebo;

  - summarization: we can have a lot of collected-and-described data and so summary measures, such as the average reductions in cholesterol level of members in the control group and in the drug group, should be determined.

Drawing conclusion from the data

- The part of Statistics concerned with the drawing of conclusions from the collected data is called **Inferential Statistics**.

  In the example of cholesterol-lowering drug, suppose that in a double-blind randomized trial the group receiving the drug has an average reduction in cholesterol level larger than the control group.

  Can we conclude that the drug is effective in reducing blood cholesterol levels?

  We must take into account the possibility that this happens by chance.

  Is this result due to the drug? Is it possible that the drug is really ineffective and that the improvement has been just a chance occurrence?

- Let's try to better understand this question of chance.

  Suppose that we want to known whether or not Head appears more easily than Tail when we throw a given coin. Suppose that we throw the coin 10 times and 7 times Head appears. Can we conclude that Head appears more easily than Tail?

  We cannot, the discrepancy between the number of Heads and the number of Tails is due to pure chance.

  The conclusion would be different if we throw the coin 50 times and 47 times Head appears.

  In this situation, to conclude whether or not Head appears more easily than Tail, the Inferential Statistics needs to quantify the chance of having at least 47 times Head over 50 throws when Head and Tail have the same chance to appear.

  The number result of this quantification is called probability. The smaller the probability, more confident we are about the fact that Head appears more easily than Tail.

As a rule, the inferential Statistics uses probabilities to draw conclusions from the collected data.

Exercise. In summer 2014, two american doctors were infected by Ebola virus in Liberia. At that time there did not exist a cure for this disease, which had a mortality rate around 50%. After they were moved in USA, they were successfully cured with the new Zmapp serum. Explain why we cannot conclude that ZMapp has been effective on them.

Population and sample

- In Statistics, we are interested in obtaining information regarding a given set of elements, called the **population**.

  Examples:

  - the set of the children beginning the primary school (in the case of the age differences in beginning of the primary school);
  - the set of the human beings (in the case of the lowering-cholesterol drug).

Often, the population is too large for examine all its members.

Examples:

- all the human beings (for example, when we test a drug);
- all the residents of a given city (for example, when we want to know the degree of satisfaction in parking car services of the residents);
- all the television sets produced in the last year by a particular manufacturer (for example, when we want to know the percentage of defective sets).

When the population is too large, we can try to learn about something regarding the population by choosing, and then by examining, an its small subset called a **sample** of the population.

- A sample must be **representative** of the population, i.e. the conclusion reached for the sample must be valid also for the population.

  Example: suppose we are interested to know the age distribution of people residing in a given city. In order to do this:

  - *we annotate the ages of the first 100 people entering the town library in a given day;*
  - *it turns out that the average age of these 100 people is 46.2 years.*

  Are we justified in concluding that this is also the average age of the entire population?

We aren't, because usually more young or old people use the library than working-age citizens.

The sample of the first 100 *persons entering the library is not representative of the population.*

A representative sample is such that all parts of the population have the same chance to be included in the sample.

This can be done in an unique way: generate the sample by picking elements from the population in a random manner.

Once a representative sample is chosen, we can use statistical inference to draw conclusions about the total population by studying the data obtained by the sample.

Exercise. A disastrous prediction of the winner in the 1936 US presidential election, in which Franklin Roosevelt clearly defeated Alfred Landon, was made by the magazine Literary Digest. A victory for Landon was predicted by the magazine. The prediction was based on the preferences of a sample of voters chosen from lists of automobile and telephone owners.

- Why the Literary Digest's prediction failed?
- Would the approach used by the Literary Digest be better today?

Exercise. For her/his master thesis, a student has to study the degree of implementation of the lean philosophy in SMEs in Pordenone province. The student prepares a questionnaire and decides to propose it to a sample of 15 enterprises. For the choice of the sample, she/he asks to the tutor professor, which prepares a list based on her/his links with the enterprises.

- Explain why this approach to the choice of the sample is not correct.
- Propose a correct approach.