Basic Statistics and Probability

Chapter 1:
Statistics, Data and Statistical Thinking

- Introduction to Statistics
- Elements of Statistics
- Statistical Procedures
- Main Types of Data
- Sampling
- Structure of Data Sets

Introduction to Statistics

Statistics is the science of data:

collection, classification, summary, organization, analysis, presentation and interpretation of the observed information.

Example 1.1: Bodyfat. Source: DASL = The Data and Story Library, `https://dasl.datadescription.com/`

Measurements of 250 men of various ages. The percent of a man's body that is fat is a matter of concern for health and fitness. But the percentage of bodyfat is difficult and expensive to measure accurately.

These data offer correct %bodyfat measurements along with a variety of easier to find measures. Can you build a model to predict the %bodyfat from other, more easily made, measurements?

Data and description at:
`https://dasl.datadescription.com/datafile/bodyfat/`

Example 1.2: Diamonds. Source: DASL

Data on raw diamonds from the internet. Price of a diamond depends on its Carat weight, color, clarity, and cut. The data are for 2690 diamonds of a variety of weights, colors, clarity, and cut. What predicts the price?

Data and description at:
`https://dasl.datadescription.com/datafile/diamonds/`

Example 1.3: Coimbra Breast Cancer. Source: UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml/`

Clinical features for 64 patients with breast cancer and 52 healthy controls.

There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis.

Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

Data and description at: `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra`

Elements of Statistics

An individual, a case or an experimental (or observational) unit is the object (person, thing, animal, event,...) that we observe to collect the data.

Example 1.3: Coimbra Breast Cancer

The individual is each of the women in the study.

Usually data are observations on a sample of individuals, chosen from a larger population. The population is the set of all individuals that we are interested in studying. The sample is a subset of the population.

Example 1.3: Coimbra Breast Cancer

The population is the whole set of women in Portugal (or in Europe, or in the world?). The sample is formed by 116 women.

When we observe every unit in a population, then we have a census of the population. This is usually feasible only for small populations. To study a large population we select a sample of it.

A variable is a characteristic or property of each unit in the population.

In each individual we can observe one or many variables. We always observe the same variables in each of the individuals of the sample, although the values of a variable have a variability or dispersion.

Example 1.2: Diamonds

In each diamond we observe the following 5 variables: prize, Carat size, color, clarity and cut.

Statistical Procedures

Descriptive statistics uses numerical and graphical methods to look for patterns in a data set, to summarize its information and to present this information in a convenient form.

Example 1.3: Coimbra Breast Cancer

Descriptive procedures would not go beyond the 116 women in the sample, describing and analyzing the information in it.

Inferential statistics uses sample data to make estimates, decisions, predictions or other generalizations about the population from which data were sampled. We use the information contained in the (smaller) sample to learn about the (larger) population.

Example 1.3: Coimbra Breast Cancer

Statistical inference would be used to infer information (from the sample) on the whole population of women, for instance, to obtain biomarkers of breast cancer.

Main Types of Data

Qualitative Variables

They are *qualities* or attributes of the individuals. They are not measured on a natural numerical scale. They are only classified into one of a group of categories.

Examples:

- Sex of an individual: male or female (categorical or nominal variable: their values are not naturally ordered).
- Degree of side effects to an oncological treatment (high, medium, low). This is an ordinal variable: its possible values can be ordered.
- Weather forecast (sunny, rainy, partially overcast, ...) in a region.
- Presence/absence of expression of a gene in an individual.

Sometimes we give a code number to each possible value of a qualitative variable, for ease of transcription and analysis, but we cannot operate with these values. For example, if the variable is Sex, we can assign number 0 to Males and number 1 to Females.

Quantitative Variables

These are measurements recorded on a naturally ocurring numerical scale. They measure something quantifiable on each individual. Quantitative variables take numerical values.

A discrete or discontinuous variable is a quantitative variable taking only a finite or numerable quantity of values.

Examples: Number of offspring in a family, number of goals of a team in a match, number of reproductions of a certain YouTube video, number of mutations in a DNA fragment.

Continuous variables can take any value in a (finite or infinite) interval of the real line, although in practice there in always a precision limit in the number of digits of its numerical expression.

Examples: Height or weight of a person (biometric measures in general), percentage of population under a poverty level in a country, temperature, price of a good.

Sampling

Once chosen the population and the variables of interest, we proceed to collect the data (sampling). Generally, original data are obtained

- from a designed experiment;
- from an observational study.

In a designed experiment, the researcher collecting the data exerts strict control over the units (people, things,. . .) in the study.

Example 1.4: Fish Diet. Source: DASL

Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer. The original study used pairs of twins, which enabled the researchers to discern that the risk of cancer for those who never ate fish was substantially greater.

Data and description at:
`https://dasl.datadescription.com/datafile/fish-diet/`

Example 1.5: Balance-Concentration. Source: OzDASL, - Australasian Data and Story Library, `www.statsci.org/data/`

How difficult is it to maintain your balance while concentrating? It is more difficult when you are older? Nine elderly (6 men and 3 women) and eight young men were subjects in an experiment. Each subject stood barefoot on a "force platform" and was asked to maintain a stable upright position and to react as quickly as possible to an unpredictable noise by pressing a hand held button. The noise came randomly and the subject concentrated on reacting as quickly as possible. The platform automatically measured how much each subject swayed in millimetres in both the forward/backward and the side-to-side directions.

Data and description at:
`www.statsci.org/data/general/balaconc.html`

Example 1.6: Cavendish. Source: Cavendish (1798)

In 1798 Cavendish estimated the density of the Earth using a torsion balance. His 29 measurements of the density, taking water density equal to 1, were

5.50	5.57	5.42	5.61	5.53	5.47	4.88	5.62	5.63	4.07
5.29	5.34	5.26	5.44	5.46	5.55	5.34	5.30	5.36	5.79
5.75	5.29	5.10	5.86	5.58	5.27	5.85	5.65	5.39	

In an observational study, the researcher observes the experimental units in their natural setting and records the variables of interest. No attempt is made to control the characteristics of the sampled individuals.

Example 1.7: Craters. Source: DASL

With the help of satellite images, lots of impact craters on Earth have been identified; now more than 180 are known. These, of course, are only a small sample of all the impacts the earth has experienced: Only 29% of earth's surface is land, and many craters have been covered or eroded away.

The data hold information about craters. Craters from the most recent 35Ma (million years) may be the more reliable data, and are suitable for analyses relating age and diameter.

Data and description at:
`https://dasl.datadescription.com/datafile/craters`

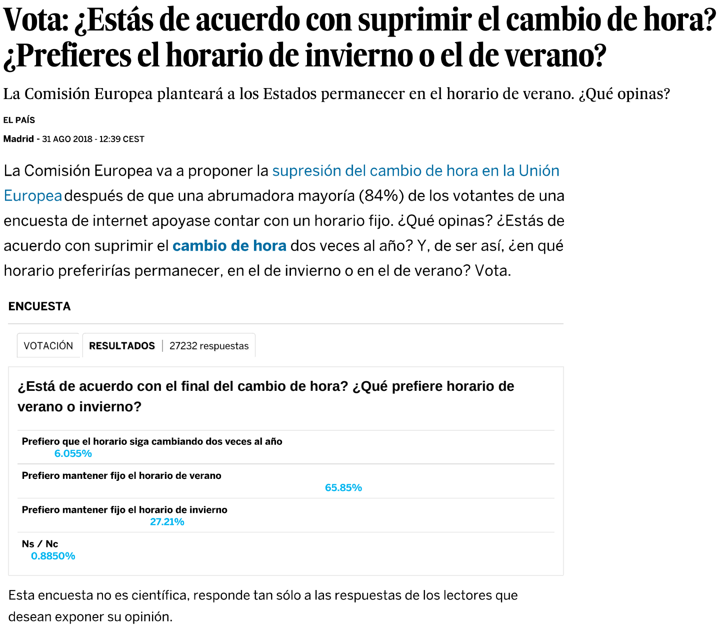
A common type of observational study is a survey, where the researcher samples a group of people, asks one or more questions, and records the responses.

A typical example of a survey is a poll, like the various ones carried out by the Pew Research Center (www.pewresearch.org).

Example 1.8: Clock changes. Source: European Commision, https://ec.europa.eu **and El País,** https://elpais.com/

Summertime arrangements in the European Union (EU) require that the clocks are changed twice per year in order to cater for the changing patterns of daylight and to take advantage of the available daylight in a given period. Following a number of requests from citizens, from the European Parliament, and from certain EU Member States, the European Commission has decided to investigate the functioning of the current EU summertime arrangements and to assess whether or not they should be changed.

Example 1.8: Clock changes



Vota: ¿Estás de acuerdo con suprimir el cambio de hora? ¿Prefieres el horario de invierno o el de verano?

La Comisión Europea planteará a los Estados permanecer en el horario de verano. ¿Qué opinas?

EL PAÍS

Madrid - 31 AGO 2018 - 12:39 CEST

La Comisión Europea va a proponer la supresión del cambio de hora en la Unión Europea después de que una abrumadora mayoría (84%) de los votantes de una encuesta de internet apoyase contar con un horario fijo. ¿Qué opinas? ¿Estás de acuerdo con suprimir el **cambio de hora** dos veces al año? Y, de ser así, ¿en qué horario preferirías permanecer, en el de invierno o en el de verano? Vota.

ENCUESTA

VOTACIÓN | **RESULTADOS** | 27232 respuestas

¿Está de acuerdo con el final del cambio de hora? ¿Qué prefiere horario de verano o invierno?

Prefiero que el horario siga cambiando dos veces al año
6.055%

Prefiero mantener fijo el horario de verano
65.85%

Prefiero mantener fijo el horario de invierno
27.21%

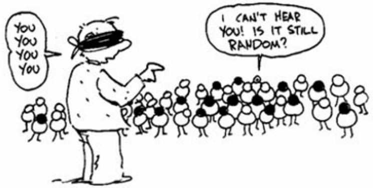
Ns / Nc
0.8850%

Esta encuesta no es científica, responde tan sólo a las respuestas de los lectores que desean exponer su opinión.

No matter which sampling method is employed, the sample should be representative of the population, that is, the sample should exhibit characteristics like those of the target population.

Random samples are the most common representative samples.

A simple random sample is a subset of individuals chosen from the population in such a way that each individual is chosen randomly and entirely by chance and has the same probability of being chosen at any stage during the sampling process.



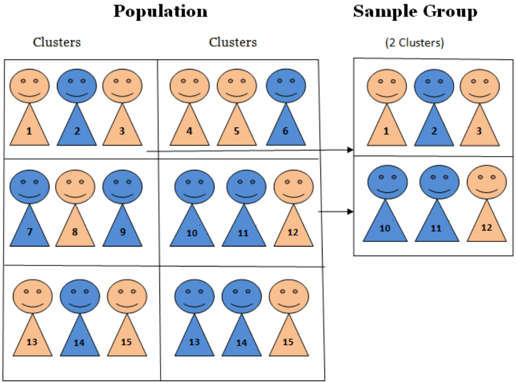
Apart from simple random sampling, there are more complex random sampling techniques.

Stratified random sampling is used when the individuals in the population can be naturally separated into two or more subpopulations, called strata, inside each of which the characteristics of individuals are more similar than across strata. This method of sampling is very useful when the population is heterogeneous.

Then random samples are obtained in each strata and these subsamples are combined to form the global sample.

Systematic sampling involves systematically selecting every *k*-th individual from a list of cases. This is typical in quality control in a manufacturing plant where items are systematically selected from the assembly line for inspection.

Sometimes we separate the population in natural groupings (clusters) of individuals. Then clusters are sampled and all the individuals from the selected clusters are observed, so a census is carried out in each sampled cluster.



Randomized response sampling is useful when the questions in a survey are likely to raise false answers. This happens typically with sensitive questions such as: have you ever tried drugs?; are you cheating on your tax payment?; are you in favour of lifetime sentences?

To circunvent the problem, each individual is presented with two questions, one being the object of the survey and the other an innocuous question. For example:

- Have you ever tried cocaine?
- Have you eaten pasta in the last 24 hours?

One of the questions is chosen at random to be truthfully answered by the person (for instance, by flipping a coin). The chosen question is unknown to the interviewer.

Structure of Data Sets

		VARIABLES				
		gender	height	weight	age	state
CASES	1	F	161	56	21	MD
	2	F	159	55	32	MO
	3	M	179	81	29	NY
	⋮	⋮	⋮	⋮	⋮	⋮
	37	M	181	80	42	TX