*Statistics is the grammar of science.*

— Karl Pearson

A large part of history of science could be summarized as an effort to translate observations of reality into precise, mathematical understanding. A record of the continuous human striving for a formulation and description of the real world in mathematical terms. To define mathematically the phenomena we find in the natural world, it is necessary to develop tools that relate the one or more relevant quantities—sometimes called *variables—*and how they relate or change depending on one another. The purpose of modelling might be, for instance, to determine the distance from the earth to the sun, to estimate the number of stars in the observable universe, or relating the number of lung cancer patients to pollution levels around smoking areas.

Colombian mathematician Luis C. Recalde marvellously summarizes the mathematical endeavour as three core tasks. For him, mathematics could be reduced to all tasks related to *count*, *measure*, and *sort*. When it comes to the description of populations, sampling, and chance, the fields of statistics and probability further develop ideas such as randomness, relationship, correlation, confidence and reproducibility, among others. Inspired by Recalde's aim to simplify, we could summarize all statistical issues as concern with *uncertainty*, or *variation* among observations.

Hence, a philosophical position often adopted is that statistics is essentially the study of uncertainty, and that the statistician's role is to assist workers in other fields who encounter uncertainty in their work. In practice, there is a restriction in that statistics is ordinarily associated with data; and it is the link between the uncertainty, or variability, in the data and that in the topic itself that has occupied statisticians. Statistics does not have the monopoly of studies of uncertainty. Probability discusses how randomness in one part of a system affects other parts.

Historically, uncertainty has been associated with games of chance and gambling. The Royal Statistical Society, together with many other statistical groups, was originally set up to just *gather and publish data*, as an attempt to reduce its uncertainty. It remains an essential part of statistical activity today, and most Governments have statistical offices whose function is the plain acquisition and presentation of statistics. It did not take long before statisticians wondered how the data might best be used, and modern *statistical inference* was born.

The mathematical formalization of decision-making is actually quite a recent development. It is usually attributed to British mathematician Frank P. Ramsey (1903–1930), who in his 1926 paper *"Truth and Probability"* [...] introduced a formal, subjective interpretation of probability, laying the groundwork for what later became expected utility theory in decision-making under uncertainty. In short, Ramsey formalized how rational agents should assign probabilities and make decisions based on personal beliefs and preferences. All starting from the apparently simple question *"how should we make decisions in the face of uncertainty?"*.

**1.1. Sampling and data types**

All statistical inquiries begin with observations and measurements, which we normally refer to as *data*. And data begins with the act of selection, or *sampling*. The natural world overflows with phenomena, offering endless opportunities for observation, but only a finite subset can ever be recorded. This distinction gives rise to two central notions: the *population*\[\mathcal{P}\], and the *sample* \[\mathcal{S}\]. By *population* we mean the complete set of all possible observations under study, normally written as

\[

\mathcal{P} = \{x\_1, x\_2, \dots, x\_N\} \; .

\]

The *sample*, on the other hand, is the finite subset actually collected. For a series of \[N\] observations \[x\_1\], \[x\_2\], ..., \[x\_N\], a sample of just \[n\] elements—less than the total, which is normally denoted by the upper case \[N\]—is defined as

\[

\mathcal{S} = \{x\_{i\_1}, x\_{i\_2}, \dots, x\_{i\_n}\}, \quad n < N \; ,

\]

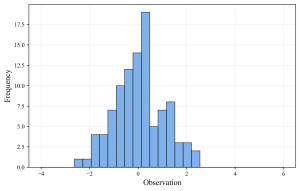
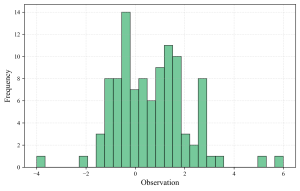
where the $i$-subscripts remind us that the sample consists of selected observations from the population, not necessarily consecutive or all of them. The population represents the ideal object of inference, while the sample is the concrete, finite evidence available to us. This distinction is far from trivial; a poorly chosen sample often misrepresents the population and may induce bias, whereas a carefully constructed one mirrors its essential features, and can be used to describe the underlying nature.

Definition of the mean value  
\[  
\bar{x} = \frac{1}{n} \sum\_{i = 1}^{n} x\_{i} \;  
\]

Population mean

\[  
\mu = \frac{1}{n} \sum\_{i = 1}^{n} x\_{i} \;  
\]

Some figures representing a histogram

[caption id="attachment\_54" align="alignleft" width="300"] Symmetric Gaussian[/caption][caption id="attachment\_55" align="alignleft" width="300"] Skewed Gaussian[/caption]