The Idea of a Probability Model for a Population

Héctor Hevia
Universidad Alberto Hurtado

Applying a phenomenological methodology, we investigate the nature and existing connections between the main objects of study that appear as a foundation in the teaching and learning of applied statistics in the first three levels of education in the Chilean educational system.

Keywords: phenomenological methodology, probability and statistics teaching, statistical and probabilistic thinking, Bruner's Cognitive Theory

INTRODUCTION

According to some authors, a phenomenological perspective in research provides "…a radical alternative to traditional understandings about what we think we can know about the world…" (Langdridge, 2007, p. 9). This type of research approach is applied in several sciences related to human behavior: for example, in nursing science (Trejo, 2012), in marketing (Green et al, 1988), and in general psychology (Langdridge, 2007).

The phenomenology considered in this article refers to that created by Edmund Husserl during his life of intense dedication to philosophical production (1889 - 1938). The phenomenological method applied to the teaching and learning of Mathematics, would allow the operating of the problems of this science, from an integrating perspective that not only involves the traditional triad constituted by the teacher, the students, and the knowledge, but also brings with it the staging of consciousness, center of all activity of knowledge of the human being. In response to Husserl's peremptory call: ***To return to the things themselves!*** a research methodology is defined with specific methods of production: *epojé*, phenomenological reduction and free variation in imagination. For a basic introduction to these methods see Chapter 2 in Langdridge (2007) and, for a more theoretical view, see San Martín (2002).

The following are the main results of a phenomenological study aimed at teaching and learning the fundamental notions of statistics: random experiments, sample space, random variables, and probability distributions, as presented in a variety of texts used in Chile, mainly in higher education.

In this work, the reader is invited to a journey where the practice of epojé, "…the process by which we try to refrain from our presuppositions, those preconceived ideas we might have about the things we investigate…" (Langdridge, 2007, p. 17), be one of his motivations.

EXPERIMENT, RANDOM EXPERIMENT AND DETERMINISTIC EXPERIMENT

An *experiment* is a replicable procedure that is carried out in reality, and from the performance of which an **observable** result is obtained.[1] In particular, we are interested in so-called *random* experiments, in which

the experimental conditions do not determine the result, i.e., *non-deterministic* experiments. For example, the experiment that consists in throwing a die at random, noting as a result the number of points that the face upward exhibits, is a random experiment. [2] However, if we modify this experiment, noting as a result the sum of the points exhibited by the four lateral faces, this latter experiment is not random but deterministic. The reason is due to the structure of the marks on a die: marks on opposite faces always add up to 7; therefore, the sum of the points on the lateral faces is equal to $(1 + 2 + 3 + 4 + 5 + 6) - 7 = 21 - 7 = 14$.

THE POPULATION OF OBSERVATIONS OF THE RESULTS OF A RANDOM EXPERIMENT

The observations of the results obtained through realizations of a certain random experiment constitute *the population of observations of the results* of that random experiment, which can be understood at the level of reality or, at a higher level of abstraction, conceptually, as a category. See Bruner (1978).[3] From a constructivist perspective, the concept "population of observations (of the results) of a random experiment" should have a relevant role in the process of teaching and learning statistics, since this is the population that we seek to model. For this purpose, a *sample* of the population is used, which is formed by a finite number *n* of these observations. The following assumption allows us to consider a sample of a size *n* sufficiently large, as a **laboratory of the population**. (It can be shown that if *n* is sufficiently large, there is no reason to object to the ability of a sample of this size to represent the population.)

THE ASSUMPTION OF INDEPENDENCE OF THE REALIZATIONS OF A RANDOM EXPERIMENT

Intuitively, we consider each random experiment as constituting an isolated, closed, and independent unit: a **monad**. Therefore, we accept that there is no possible effect on the result of the random experiment that does not come from the experimental conditions and that there is no effect resulting from the performance of the experiment; we accept, then, that there is no other type of causality involved in the experiment and that, therefore, the repetition of the random experiment produces *independent results*. We call, this a priori conception of the random experiment, *the assumption of independence of the realizations of a random experiment*.

THE RESULTS OF A RANDOM EXPERIMENT

The definition of what are the results of a random experiment being studied is entirely in the hands of those who study such an experiment.[4] These results, called *simple*, can have two natures. We call, results of a random experiment that can be observed through performing the experiment, *phenomenal results*[5]. On the other hand, we call, those results whose existence is guaranteed by accepting certain assumptions, *theoretical results* of the random experiment. [6] For example, if the random experiment consists of throwing a die at random one hundred times, noting as a result the sum of the hundred numbers obtained, the result "100" which is obtained when in each of the hundred throws "1" is obtained, could be catalogued as theoretical, but possibly not phenomenal.

The results of a random experiment could be established using a convenient coding. In this case, it is recommended that the coding of the results be integrated into the definition of the experiment; ambiguous experiments are unsuitable for learning. Let us look at two examples.

Example 1

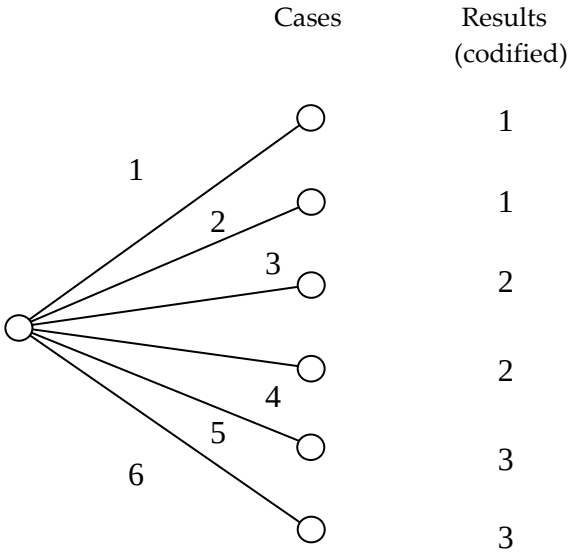
Suppose that the random experiment consists of throwing a die at random, scoring "1", "2", "3"; depending on whether the number of points observed is "1" or "2"; "3" or "4"; "5" or "6"; respectively. Note how the defined coding makes it possible to decide whether two throws produce the same result; that is, when the results obtained are considered "equal". Figure 1 shows an iconic representation of this random

experiment in which the lines between nodes correspond to each of the possible productions or *cases* of the experiment.

Example 2

Suppose that two dice are thrown randomly and it is only of interest whether the number obtained in each throw is odd or even. The experiment is so imprecise that there are several alternatives for recording a result. Let us choose, as coding for the results, the letter *o* and the letter *e* to indicate that a die threw an odd or an even number, respectively. Furthermore, let us accept that each letter will preserve the identity of the die that produces its observation; for example, by noting the result of one of the dice first and the result of the other die second. Then, there are exactly four possible results: *ee*, *eo*, *oe*, *oo*. Note that, if we ignore the identities of the dice from which the observations come, then, there are exactly three possible results for this experiment: *ee*, *eo* = *oe*, *oo*. In Figure 2, an iconic representation of this experiment is shown indicating cases and their respective results.

FIGURE 1
AN ICONIC REPRESENTATION OF CASES AND RESULTS OF EXPERIMENT IN EXAMPLE 1



THE ASSUMPTION OF EQUIPROBABILITY OF THE CASES OF AN EXPERIMENT

If the mode of production of the results of the random experiment is accepted as known and this mode of production determines a **finite** number of possible cases to be described, then, under *the assumption of equal occurrence of the cases* (or *equiprobability of cases*) it is possible to assign probability to a simple result *R* of the random experiment through the well-known Laplace's Rule; which we highlight as follows.

$$P(R) = \frac{\text{number of cases in favor of } R}{\text{total number of cases}}$$

This way of assigning probabilities to the simple results of a random experiment corresponds to the so-called *classical method* of assigning probabilities.

In the first example and proceeding in this way, we have that two cases, of the total of the six cases, are favorable to each possible result, assigning, therefore, probability $\frac{1}{3}$ to each of the three results. In the second

example, there are 36 cases or productions of the experiment; of which, 9 are favorable to *ee*, 9 favorable to *eo*, 9 favorable to *oe* and 9 favorable to *oo*; assigning, therefore, probability $\frac{1}{4}$ to each result. Note that if in this second example the reference to the identity of the die originating the observation is ignored, i.e., noting the results as unordered pairs, there are 9, 18, 9 cases favorable to *ee*, *eo* = *oe*, *oo*, respectively; therefore, in this situation the probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, respectively, are assigned.

SAMPLE SPACE OF A RANDOM EXPERIMENT AND ITS FAMILY OF EVENTS

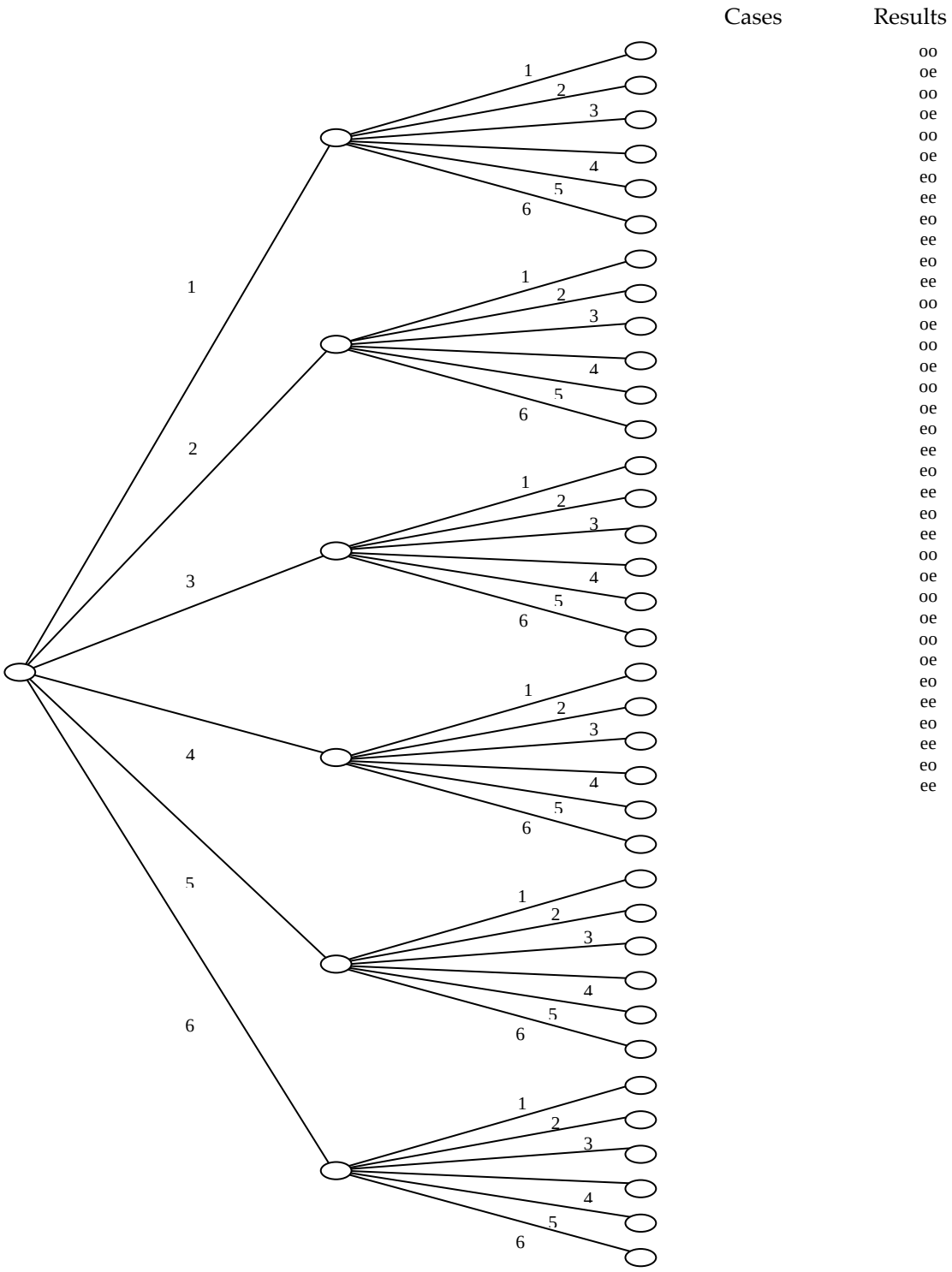
Let *E* be a random experiment. The *sample space S* of the experiment is defined as the **set** whose elements are all the simple results that, **in theory**, can be obtained from the experiment *E*.[7] Note that a convenient representation of the results of the experiment, allows to describe the elements of *S* through terms or expressions whose meanings could maintain a degree of relation with the random experiment under study. From this perspective, the sample space of a random experiment constitutes a first-generation **mathematical model**[8], a **mathematical object** projected on a context of reality; therefore, imbued with possible meanings[9]. For example, the sample spaces of the three examples presented above are: $S = \{1,2,3\}$, $S' = \{ee, eo, oe, oo\}$ and $S'' = \{ee, eo, oo\}$; respectively.

The elements of a sample space, i.e., the simple results of the random experiment, are also called *simple events*. For example, *oo* is a simple event for the last two sample spaces.

In general, an *event* is a **statement** that alludes to, or refers to, the results of a random experiment. For example, the statement "the result is an even number" is an event for any random experiment that yields an integer number as a result. But these statements, which constitute events of a random experiment, must be feasible to be verified for each simple result of the experiment.[10] Whatever such statement is stated in relation to the theoretical results of a random experiment, it will define, by understanding, a **subset of the sample space** of the experiment.

Understanding events as statements in natural language referring to the simple results of a random experiment, it is possible to enrich the learning of probabilities by making it possible to integrate several sample spaces of interest in which the same event is relevant, but whose respective subsets could be different. Consider, for example, the event *E*: "one of the dice throws an even number". Then, $E' = \{ee, eo, oe\}$ and $E'' = \{ee, eo\}$ are representations of *E* in the sample spaces S' and S'', respectively. This flexibility achieved through this notion of event, could have interesting implications in the teaching and learning of some applications of Bayes' Theorem, in which probabilities of different origin are combined.

FIGURE 2
AN ICONIC REPRESENTATION OF CASES AND RESULTS OF EXPERIMENT IN EXAMPLE 2



THE REPRESENTATION OF A RANDOM NUMBER: RANDOM VARIABLE

In what follows, we consider random experiments the results of which are **numbers**. Undoubtedly, through these random experiments, the so-called *random numbers* become present.[11] Suppose E is a random experiment whose simple results are numbers. We use a capital letter, say X, to represent the *random number* that is present through the experiment E and that is susceptible to be observed through a realization of the experiment E. The term X is called a *random variable*. On the other hand, the number that is observed through the performance of the random experiment E, is denoted by x. We call this definition the *phenomenological definition* of a random variable.

Note that a random variable X, understood in this way, represents a random number and that these random numbers are feasible to be recognized intuitively, thanks to the presence of a certain random experiment. This definition makes it possible to apprehend the concept to which a random variable refers, i.e., "random number", in the same way in which numbers are apprehended: as phenomena that are directly recognizable in reality.[12]

In fact, sooner rather than later, the authors of higher education textbooks, are inclined to this conception of random variable since the other existing alternative definition, here called *functional definition*, which conceives a random variable as a function of the sample space of a random experiment to the set of real numbers, becomes untenable when the random variables being studied are continuous. Others, such as Rice (2007), Durrett (2009), Lind et al. (2015)[13] and Ross (2002); have opted, practically from the beginning, for the phenomenological definition, pointing out, for example, "a random variable is essentially a random number"[14].

If we opt for an understanding of the concept of random variable through a phenomenological definition such as the one already mentioned, it is necessary to previously consolidate the concept of random experiment, which should be an object of study with this new objective in mind: the formation of the concept of random number through learning experiences involving random experiments whose simple results are numbers. The previous manipulation of random experiments at appropriate school levels would allow the incorporation of meanings for random numbers, which would be directly linked to the students' experiences. This would also contribute to the formation of iconic representations for more general random experiments, which have a prominent role in the construction of probability concepts. (On the types of representations, see Bruner (1964).) Ideally, it could be feasible for these objectives to be achieved in the years prior to high school. Following this line of research, a preliminary investigation is being developed with the objective of determining the status of the object "random experiment" in Chilean school textbooks (Figueroa & Hevia, in preparation).

The following is a brief digression on the teaching of the concept of random variable when the so-called "functional definition" is used. The usual, in statistics textbooks, whatever the level of studies, basic, intermediate, higher (undergraduate), consists of defining "random variable" as a mathematical function whose domain is the sample space S of a random experiment and whose path is the set of real numbers. From the point of view of learning, there are more than one reasons that would make this definition inadvisable. First, whether the random experiment delivers numerical results or not, such a definition is of dubious utility since it would not go beyond being a mere encoding of the results of the experiment. Such a coding could be added as a further specification in the definition of the random experiment under study, allowing the reference to the mathematical object (the function), to be omitted. But there is a second reason that makes the definition of random variable as a mathematical function not advisable: the use of the mathematical object function to define the concept of random variable gives greater abstraction to this concept, distancing it from reality in the sense explained in Hevia (2021).

Let us return to the concept of random variable as the representation of a random number. A relevant characteristic of a random variable is that it has the character of a unique and, at the same time, multiple representation[15]. For this reason, the algebra of random variables must have its own considerations. For example, in the die experiment, the random variable X is the number of dots on the top face. Therefore, the sum of the numbers obtained in two (independent) realizations of the experiment must be represented by differentiating between these two random numbers, writing, for example, $X + Y$ or $X_1 + X_2$.[16] In the latter

case, we preserve the letter X to indicate that we are dealing with results of the same experiment E, but we differentiate by means of subscripts to indicate that they correspond to results coming from two different realizations of the experiment. Note that $T = X_1 + X_2$ is a new random variable, which could take any integer value from 2 to 12.

The introduced notation using subscripts is convenient for representing and theoretically manipulating a sample (or part) of the population; in particular, for representing *independent and identically distributed samples* of size n of a population X (in short, an i.i.d. sample of a population X). This type of sample plays an important role in estimation theory and in statistical inference techniques. Thus, a sample of size n of a population X could be represented by the system of random variables $X_1, X_2, \dots, X_n$.[17]

Note, in the previous paragraph, the naturalness of the use of the random variable X to refer to the population of observations of the random experiment where the random number X is present.

THE ASSUMPTION OF THE EXISTENCE OF A PROBABILITY MODEL FOR THE POPULATION OF OBSERVATIONS OF THE RESULTS OF A RANDOM EXPERIMENT E WHOSE SAMPLE SPACE S IS FINITE

It seems practical to abbreviate "population of observations of the results of an experiment" to "population of observations of an experiment", which we will do hereafter.

Given a random experiment E, it makes sense to assign a single mathematical object to the population of observations of E, whose role is to establish a law for the probabilities that are likely to be present in the observations of experiment E. This mathematical object serves the role of a *probability model* for the population of observations of the experiment E by establishing a reference standard for the relative frequencies appearing in a sample of the population of observations of E.

In particular, if the set of theoretical results of the random experiment is finite[18], a mathematical function that assigns probability to each of these possible results, called the *probability mass function (pmf)*, is used as a model. However, there is no reason to affirm that the population of observations of any random experiment whose sample space is finite has such a model, not even in the case of more general experiments. We are in the presence of a new assumption: *the assumption of the existence of a probability model* for the population of observations of a random experiment E, whose sample space S is finite.

Let E be a random experiment whose sample space S is finite. Let X be the random number that is present in the random experiment E and suppose that there exists a ***measure or probability of occurrence*** for each theoretical result x of the experiment E, denoted by $P(X = x)$[19]. Then,

$$f(x) = P(X = x); \text{ for every } x \in S,$$

is called a *probability mass function* of X.[20] This probability mass function $f(x)$ is interpreted as a probability model for the population of observations of the experiment E. Even more briefly, one could say that $f(x)$ is a probability model for the population X.

For example, if E is the die experiment, under the assumption of equiprobability of the cases, one obtains.

$$f(x) = \frac{1}{6}, \qquad \forall\, x = 1, 2, 3, 4, 5, 6$$

This probability mass function can be interpreted as a probability model for the population X of the die experiment.

In the previous example, the probability model has been derived under the assumption that an honest die does not privilege the occurrence of any of the six outcomes and that, therefore, equiprobability is an inherent characteristic of this type of die. However, nothing guarantees that the inferred model is valid for

the population of observations under study. This will inevitably lead to the need for **validity tests** for the inferred model.

CONCLUDING REMARKS

Since a table of values gives probabilities without recourse to the mathematical object function, it would be possible to assimilate the concept of a population model, dispensing with the mathematical object function; at least, in the first years of study. For example, see Lind (2015).

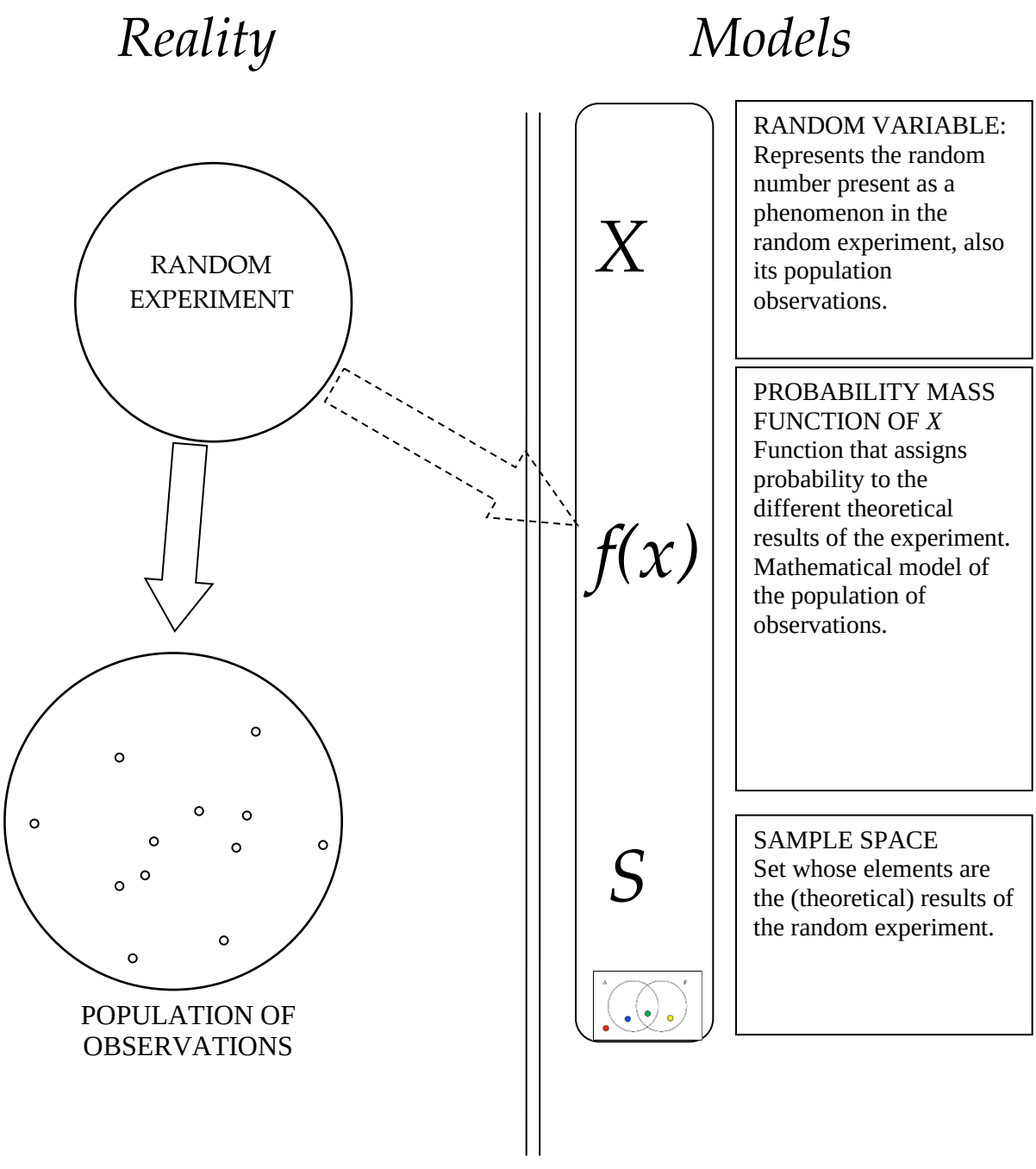
If the probability distribution function, as in general a pmf is called, is not presented as a mathematical model of the population of observations, then the statistical concept of "fit of a model to the data" is meaningless.

Let us summarize what we have seen, representing: ***random experiment***, ***result of a random experiment***, ***observation (of a result of a random experiment) and population of observations***, ***sample space***, ***simple events, and compound (non-simple) events***, ***random variable***, ***probability distribution*** in a diagram that allows visualizing or at least perceiving the existing relationships among them (see Figure 3). Three of these objects: **random experiment**, **population of observations** and **sample space**, appear as foundational concepts for the construction of statistical and probabilistic thinking. Note that, in their natures, these foundational concepts are a process, a category, a mathematical model (involving a mathematical object).

From the figure below, it would seem advisable that students are able to recognize these objects of knowledge, as a necessary preparation for the beginning of their studies in statistics[21]; otherwise, it is possible that inevitable obstacles in the development of students' statistical thinking take place.

In Figure 3, the dotted arrow indicates that, if the mode of production of the results of the random experiment is fully known, the probability model of the population could be constructed exclusively through purely theoretical considerations. However, it could never be tested as such, but validated. In the first instance, these validity tests could be based mainly on the so-called (weak) Law of large numbers. See Ross (2002). On the other hand, the continuous arrow would establish an order of precedence between random experiment, observations, and the population of observations.

FIGURE 3
IF THE RESULTS OF THE RANDOM EXPERIMENT ARE NOT NUMBERS, THEN THE RANDOM VARIABLE X IS NOT DEFINED



ENDNOTES

1. Experiments whose results are not observable are not of interest. For example, it is understood that a die enclosed in an airtight box with non-translucent walls, wide enough for the die to rotate and move arbitrarily, produces random results each time the airtight box is shaken and then stopped; however, this experiment is not of interest since its results are not observable. In fact, opening the box would violate the original design

of the experiment and, consequently, little or nothing could be said about the results of the originally designed experiment.

2. We will call this experiment, "the die experiment".
3. At the level of reality, the population of observations could be constituted by the observations already made; but it is also possible to abstract from reality and conceive the population as being constituted by all those observations of the results of the random experiment that result from an execution of the experiment. In this way, the population becomes conceptual and infinite; infinite in the sense that it can be as large in number as one wants it to be.
4. There will always be a degree of reduction of reality in establishing what is to be understood as a result of a certain experiment: Whenever something is defined, there is also something that is lost.
5. Phenomenal: pertaining or relating to the phenomenon as an appearance or manifestation of something.
6. Keep in mind that what is assumed is not always true.
7. The notion of sample space comes from R. von Mises; as mentioned by Feller (1950).
8. In a first-generation mathematical model (Gårding,1977), the mathematical objects of the model correspond to objects present in a context of reality, which allows meaning to be given to the abstract objects that make up the model (Guzmán & Hevia, 2002).
9. According to Duval (2001), mathematical objects, being abstract, are only attainable through representations.
10. Quoting Feller (1950), p 14: "…It makes complete sense to talk about an event A only when it is clear for each outcome of the experiment whether event A has or has not occurred."
11. It should be noted that a random number is not a number. On the other hand, the observation of a random number does result in the recording of a number.
12. This random variable teaching proposal was presented for the first time in the micro course Phenomenological Didactics of Statistical Modeling of Data (Didáctica Fenomenológica de la Modelación Estadística de los Datos) given by the author at the First International Congress of Statistics, Universidad de Trujillo, Peru; 2013.
13. In particular, Lind et al. (2015), p. 157, identify "random experiment" with "random variable", expressing: "In any random experiment, the results are presented at random; thus, this is called a random variable"
14. Rice (2007), p. 35, begins the chapter dedicated to random variables by stating: "A random variable is essentially a random number."
15. It is important to take into consideration that a random variable X represents a single observation of a result of the experiment; hence the uniqueness of the representation. Simultaneously, this observation can refer to any of the values of the variety of results produced by the experiment; hence the multiplicity of the representation.
16. Note the naturalness of the presence of several dimensions, already at the beginning of a study on random numbers and their representations.
17. The samples defined in point 3, under the assumption of independence, point 4, are i.i.d. samples.
18. Or numerable infinity.
19. This assumption accepts the existence of a probability model for the population of observations of E.
20. It is understood that these probabilities satisfy the basic laws; for example, those established by Kolmogorov's axioms; although simplified, since it has been accepted that S is finite.
21. It seems appropriate to establish as an objective of Basic Education in the thematic axis "Statistics and probabilities", that these three fundamental concepts should be part of the knowledge achieved by students in the first years of study and prior to any other study of statistics.

REFERENCES

Bruner, J.S. (1964). The course of cognitive growth. *American Psychologist*, pp. 1–15.

Bruner, J.S., Goodnaw, J.J., & Austin, G.A. (2017). *A Study of Thinking*. Routledge.

Durrett, R. (2009). *Elementary Probability for Applications*. Cambridge University Press.

Duval, R. (2001). Un tema crucial en la educación matemática: La habilidad para cambiar el registro de representación. *La Gaceta de la RSME*, *9*(1), 143–168.

Feller, W. (1950). *An Introduction to Probability Theory and Its Applications* (3rd Edition). John Wiley & Sons, Inc.

Figueroa, R., & Hevia, H. (in preparation). See, Figueroa, R. (2022). *Experimento aleatorio, su estatus en los textos de estudios en Chile*. Final work for the Master's degree in Didactics of Mathematics, Facultad de Educación, Universidad Alberto Hurtado, Chile.

Gårding, L. (1977). *Encounter with Mathematics*. Springer-Verlag, New York Inc.

Green, P.E., Tull, D.S., Albaum, G. (1988). *Research for Marketing Decisions* (5th Edition). Prentice Hall.

Guzmán, I., & Hevia, H. (2002). *Modelos Matemáticos y su Incidencia en el Aprendizaje de las Matemáticas.* Published in Informe Técnico CIDIC 2, Universidad Adolfo Ibáñez, Chile (2008).

Hevia, H. (2021, April–June). Incidencia de los Paradigmas de la Matemática en la Enseñanza y Aprendizaje de la Estadística. *South Florida Journal of Development, 2*(2). DOI:10.46932/sfjdv2n2-094

Langdridge, D. (2007). *Phenomenological Psychology. Theory, Research and Method*. Pearson Prentice Hall.

Lind, D.A., Marchal, W.G., & Wathen, S.A. (2015). *Statistical Techniques in Business and Economics* (15th Edition). McGraw-Hill.

Rice, J.A. (2007). *Mathematical Statistics and Data Analysis* (3rd Edition). Thomson.

Ross, S. (2002). *Introduction to Probability and Statistics for Engineers and Scientists* (2nd Edition). McGraw-Hill.

San Martín, J. (2002) *La estructura del Método Fenomenológico*. Universidad Nacional de Educación a Distancia, Madrid.

Trejo, F. (2012). Fenomenología como método de investigación: Una opción para el profesional de enfermería. *Revista de Enfermería Neurológica (Mex)*, *11*(2), 98–101.