CONNECTING MATHEMATICAL AND STATISTICAL TEACHING: THE ROLE OF STATISTICAL VARIABLES

Guido del Pino and Ana María Araneda
Pontificia Universidad Católica de Chile
gdelpino@mat.puc.cl

Teaching statistics is unappealing to most high school teachers, in part due to their deficient university training in this field and to the fact that people with a strong mathematical preparation tend to look down on areas that use only elementary mathematics. To lessen this problem this paper establishes some connections between descriptive statistics, probability, and mathematics, comparing statistical variables, random variables, and functions. Several arguments are given to justify that in statistical applications, it is not the function representing the variable that matters, but the key element is the induced frequency distribution, which is important for the teaching probability theory. Some pedagogical ideas are given for teaching continuous variables, a topic that poses subtle mathematical issues, like the meaning of a real number, of continuous functions, and of infinite domains, and that the exact value of a continuous variable is not observable.

INTRODUCTION

Teaching statistical ideas to people with a good mathematical background, but a weak statistical background, such as future mathematics teachers, is not easy. They tend to look down on areas that use only elementary mathematics, which is perhaps one reason why textbooks addressed to engineers and scientists put more emphasis on the mathematical aspects (Devore, 2011; Montgomery, 2010; Walpole, Myers & Myers, 2011). (Usiskin, 2014) claims that statistics can be made more attractive by establishing deeper connections with mathematics, and that these connections are useful for a better understanding of both disciplines. This paper is written along these lines, by formalizing the simple concept of statistical variable X as a function f, where the qualifier "statistical" is used to make a distinction from "random". The idea clearly comes from the definition of random variables, but we claim that it is better to teach statistical variables before teaching random variables.

We do not find anything wrong with defining a variable as a characteristic of individuals, which may change from one to another. One of our objectives is to provide concrete and relevant examples of the abstract definition of function. High school students first encounter this topic in algebraic expressions, such as $x^2 - 2x + 3$, which translates into a function through $y = f(x) = x^2 - 2x + 3$. In modern mathematics (Jacobs, 1979) a function is "any rule that associates with every element s of a set D (called the domain) a unique element (its value x = f(s)) in some set C (called the codomain). The typical representation is a Venn diagram with sets and arrows, but for a finite D it is also useful to use an m x 2 table, where m is the number of elements of D, whose elements do not have an inherent unique ordering. The rows of the table have the form (s, f(s)), but their order is defined by one of the m! possible enumerations of the population elements. In any case, these m! tables are equivalent in the sense that they represent the same function. In statistics, the domain can be either the population or the sample, and a variable X assigns to s the value f(s) of some function f. A similar discussion arises for random variables, but we postpone it. We claim that the statistical framework leads to a functional representation, which should appeal to mathematicians. A further point is that the type of a variable is essentially determined by the codomain of the function.

In this paper, we explore in detail the use of functions to represent statistical variables and, in particular, to help discussing their types. We carefully analyze the mathematical difficulties posed by continuous variables. We present several arguments to justify that the only important aspect of a statistical variable is its frequency distribution, and we emphasize the role of permutation invariance. We argue in favor of dropping the standard restriction to real valued functions. We propose several classroom activities that exploit the ideas presented in the paper. We end with a discussion and some closing remarks.

STATISTICAL VARIABLES AND FUNCTIONS

In: M.A. Sorto (Ed.), *Advances in statistics education: developments, experiences and assessments. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),* July 2015, Rio de Janeiro, Brazil. ©2015 ISI/IASE iase-web.org/Conference_Proceedings.php

Real variables are sometimes described as "numbers that change", but there is logical flaw here since, for example, number 2 cannot change. Statistics makes the point clearer, by allowing *x* to be the family size of John and *y* to be the family size of Mary, and *y* may be differ from *x*. This contrast means that the variability is between persons, and not within persons. Since every individual has a family size, it is natural to view family size as a function. In the physical world, everything is finite (even the set of all atoms in the universe), so that nothing is really lost by restricting ourselves to finite populations and finite set of potential values of a variable. However, if X is number of accidents, the codomain could be taken as the set of nonnegative integers, while if Y denotes height, the codomain is often chosen as a real interval. Strictly speaking, these infinite sets are mathematical constructs that live only on our minds, but are convenient to tackle some practical problems. Though the concept height exists in physics, for statistical applications we apply it either in the population or in the sample, but the different domains imply that the statistical variables are not represented by the same function. Let f and g represent height at the population and the sample, respectively. Then g(s) = f(s) for all s in the sample, which, in mathematical terms means that g is a *restriction* of f and that f is an *extension* of g, this second concept being somewhat related to statistical inference. Types of variables are discussed both in descriptive statistics and in probability, but we concentrate on the first in this section. (Moore, 2010) says that "*A categorical variable places an individual into one of several groups or categories. A quantitative variable takes numerical values for which arithmetic operations such as adding and averaging make sense. The values of a quantitative variable are usually recorded in a unit of measurement such as seconds or kilograms*". While categorical and numeric variables are always discussed, the distinction between discrete and continuous variables is sometimes omitted. (Utts & Heckard, 2007) discuss initially all types of variables, but they postpone the mathematical definition of discrete and continuous variables to the probability chapter, where they say that "a discrete variable has a finite or countable set of real values, while the set of values of a continuous variable is a real interval". A subtle mathematical difficulty is that a variable whose set of possible values is an interval of rational numbers would be discrete, but in practice one would always treat it as continuous. In the functional framework variable types are determined by the nature of the codomain:

I. Categorical variables (also known as qualitative or attributes): The codomain is finite and its elements are called categories.
Nominal : There is no meaningful order, even if integer codes are used.
Ordinal : The categories have a natural order, being often represented by consecutive integer codes.

II. Numeric variables (also known as quantitative): The codomain is a set of real numbers.
Discrete : The codomain is finite or countable.
Continuous : The codomain is not necessarily countable.

Remarks: 1) Discrete variables need not be counts; 2) In the sequence nominal-ordinal- discrete-continuous, we assume that each variable type includes the preceding ones. A mathematical treatment of continuous variables is not easy. We know that any real number can be extremely well approximated by numbers with a large but finite number of decimal places. On the other hand, in practical applications we can replace an unbounded set by a bounded one. By definition, a continuous variable X can then be approximated by a finite variable Y. Though this is quite enough for statistical applications, physical variables, like the temperature in a room, are treated as exact, and they are allowed to vary continuously in a subset W of the Euclidean space. Of course, we can measure the variable at the points of a finite set S that will be taken as the population. The trouble here is that sample values would then be exact real numbers, which are not observable due to the unavoidable measurement errors. A mathematical solution is to use a rounding operation R, such that when it is applied to a real number x produces one with a finite number k of decimal places. The larger k is, the better will be the approximation, but observability forces k to be relatively small. The new function h, given by h(s) = R(f(s)), defines a discrete variable.

FREQUENCY DISTRIBUTIONS AND THEIR REPRESENTATIONS

Our emphasis on functions does not imply that all of their aspects are relevant for statistical analysis. If a table represents a function, we can construct two main tables. Table 1 sorts the rows according to the first column (the identifiers) and Table 2 sorts the rows according to the second column (the values of each variable), which are assumed ti be ordered. For a categorical variable one needs to impose some arbitrary order in the categories, for instance, alphabetical. For instance, in a population of 40 men and 60 women, the second column in Table 2 consists, for instance, of 40 men followed by 60 women, and this represents the variable "gender".

In descriptive statistics, the most important case is that of a sample variable, such as height. The second column contains the heights, ordered from smallest to largest, which yields the order statistics. Though the first column is crucial to define the function, in most statistical analyses part of this information is irrelevant. (del Pino, 2014) shows that the precise mathematical condition needed to examine this issues is that of *exchangeability*, which, in mathematical terms, means invariance under permutations. If S is the population with size 8 whose elements are S= {1,2,2, 4,5,7,7,7}, and sampling is done without replacement the frequency of each element in the sample is at most equal to the corresponding frequency in the population. Sampling with replacement greatly simplifies the analyses, since these bounds are irrelevant, and the sample size is arbitrary. In the previous example, a possible ordered sequence of sample values is
{4, 4, 3, 5, 2, 2, 7, 4, 2}, which can be arranged as a second column in the table. Note that the population elements associated to each value are not given here. Since the order is irrelevant for most statistical analysis, from a combinatorial viewpoint we are dealing with combinations with repetitions. In statistical terms we are dealing with combinations with repetitions. In any case, the lack of the first column in the population makes it impossible to compute the absolute frequency distribution, that is, a function which to each elements associates its number of repetitions (its frequency). Letting g(x) denote the frequency of x, the function g is a representation of the frequency distribution. In the example g(1) = 0, g(2) = 2, g(3) = 1, g(4) = 3, g(5) = 1, and g(6) = 0, and g(7) = 1. For the gender example, g(man) = 40, g(woman) = 60. (del Pino, 2014) proves that, under an exchangeability condition, the frequency distribution is a maximal invariant statistic. This means that it not only contains all of the relevant information, but it also gets rid of all the irrelevant information inherent in order. It is a mathematical result that all maximal invariant statistics are in one to one correspondence, so that we call each one an alternative representation. Much of descriptive statistics can be stated in terms of obtaining equivalent representations of the sample frequency distribution. We make this explicit, separately for each variable type:

Nominal: Frequency table and bar graph. The order of the categories need not be relevant.
Ordinal: Nominal variable whose categories have a natural order. The table and the graph must respect this order. Cumulative frequency distributions make sense.
Numeric: An ordinal variable, where the order is given by that of the real numbers, and arithmetic operations on the values are meaningful. Two useful equivalent representations are the vector of order statistics and the dotplot. On the other hand, the graph of the cumulative frequency distribution is piecewise constant, with jumps at the sample values. For a continuous variable, there will typically be few repetitions in the sample, particularly if the value has been recorded with many decimal places. In the limit there will be no ties, so that all frequencies will be equal to 1, so that the only relevant information is then the unordered collection of observations.

Population frequencies are normally much larger than sample frequencies, so that a direct comparison is not meaningful. A solution is to only compare the *relative* frequency distribution of the sample with that of the population. It has several interesting properties

- To understand a table of absolute frequencies, percentages are more convenient than the absolute frequencies.
- In graphical terms the absolute and relative frequency distributions have the same *shape.* (only the ordinate scale changes).
- Ratios of relative frequencies are identical to ratios of the absolute distributions.

- In random samples, the relative frequency distribution of the population and the sample will be close for large samples.

RANDOM VARIABLES

In probability, a random variable X is defined as a real valued function defined on a sample space S, together with some probability measure. We claim that this definition defies intuition, with the ensuing difficulties for students, who must simultaneously cope with the probability axioms and with the mathematical definition of function. On the other hand, probability functions and distributions are introduced at a second stage. To illustrate this difficulty, let X be the number of accidents in a day, and assume it follows (approximately) a Poisson distribution. It is hard to discover the function f in this setting. The typical mathematical solution is (a) choose S as the set of all nonnegative integers (b) assign to each s in S its probability using the Poisson formula (c) take f to be the identity function. This three-stage approach is a rather cumbersome mathematical argument that does not add anything useful. To shed some light on the problem, consider a large finite population and a variable X with a relative frequency distribution u. It is known that choosing an element at random from this population produces a random variable Y, whose probability function coincides with u. The connection with descriptive statistics suggests that it is not the function f representing X that matters, but rather it is rather its probability function, from which it is impossible to find the function f. On the other hand, the standard definition assumes a random variable X to be real valued, a limitation that we consider unnecessary and obscures the connection with statistical variables. Consider again the experiment of choosing at random a person from a population of size 100. Natural choices are a sample space S = {1, 2, …, 100}, and X(s) = gender of s, which is not real valued. The probability function of the variable gender exists and it is p(man) = 0.4 and p(woman) = 0.6. Something similar occurs with the variable "opinion" about a future law. Something similar occurs with ordinal variables, such as those measured in a Likert scale: *strongly disagrees, disagrees, neutral, agrees*, and *strongly agrees*. In statistical inference, a technical option is to use five indicator variables, an approach useful to analyze the data, but it does not help understanding the model.

ACTIVITIES

We present activities that exploit the ideas presented in this paper for statistics, mathematics and probability classes. Activities on statistics and mathematics are intended be used directly in a school classroom. Activities in probability are more demanding, and address to statistics instructors at a university.

Students in an elementary statistics course: The central problem in these activities is to find different and convenient ways of describing the function f representing a given variable, Table 1, ordered according to individuals, and Table 2, ordered according to the values in the function, in the text. The goal of the proposed activities is to make clear the role of the mathematical concepts. They also try to make clear the central role of frequency distributions. Although the activities refer to a population, recall that the same discussion applies without any change to samples.

Activity 1: Functions, tables, and frequency distributions.
a. Consider the population formed by the names of 10 individuals and the ordinal variable representing their opinions, with five categories (coded as 1, 2, 3, 4, 5). (i) Construct function f and represent it using an arrow diagram. (ii) Construct three different tables representing the same function f. (iii) Comment on the feasibility of drawing an arrow diagram when there are 100 individuals in the population.
b. (i) Construct a questionnaire with 4 questions, to be applied to all the N students in your classroom (say 40), making sure that all types of variables are present. (ii) Order the student names in alphabetical order, and put them in the first column of a 40 x 5 table. (iii) Note that to each name there is associated a row array (x, y, z, t). Although this array is relevant for multivariate analyses, we deal only with one variable (column) at a time, that is, with N x 2 tables.
c. Code the names, keeping a "dictionary" to understand the codes you used. (i) Argue why no relevant information is lost if somebody destroys the dictionary. (ii) Can we drop the first column

in the table? (iii) If for confidentiality reasons we cannot inform about the names: Does one lose any relevant information? (iv) Can we still obtain the frequency distribution?
d. Standard statistical analyses never use the names of the subjects. Convince yourself that this practice implies that we only need the frequency distribution, and check this claim explicitly for the mean, the median and the standard deviation.

Activity 2: A categorical variable (*opinion*) Consider Table 1, ordered by the column of names and Table 2, ordered by the data value arranged in increasing order. Discuss the relative advantages of Table 1 vs Table 2. For a large population, ordering a column can be rather expensive from a computational point of view. Find a better strategy for the example of the Likert scale. (think of a tally diagram).

Activity 3: Counts (*family size*).
a. Construct a data set for variable representing counts, and check that the discussion in Activity 2 still applies.
b. Make a dotplot of the data, compute the order statistics, and draw the empirical cumulative distribution function.
c. Check that we can construct the frequency distribution from any of these representations, and conversely.

Activity 4: Continuous variables (*height*).
a. Repeat Activity 3, part b., (i) using the original data and (ii) after first sorting the data in increasing order.
b. Compare the relative effort to perform the computations in (i) and (ii).
c. Consider a large sample of 1000 observations of height measured very accurately. (i) Argue why the typically cumulative distribution functions has 1000 jumps of size 1 (ii) If (i) holds, check that the jumps in the relative cumulative distribution function have size 1/N. (iii) What can you say about approximating the graph by a smooth curve?

Students in a mathematics course:
Activity 5:
a. Take any function with a finite domain of size 20, and make up a story that represents this function by choosing a population and a variable.
b. Check that in the "arrow graph" representing the function, the type of variable is irrelevant.
c. Represent the function by three alternative 20 x 2 tables, and think how these constructions relate to permutation invariance. (iii) Find the frequency distribution.
d. To understand combinations with repetitions (unordered samples with repetitions), make up some data, and find a story to interpret them in the context of choosing elements from a sample.
e. What is the frequency distribution?

Students in a probability course: This activity provides a simple way to understand the topic of functions of random variables and their probability distributions. Even if random variables are real valued, we can consider them as categorical, and this reduced view is what we will assume**.**
Activity 6:
a. Take a finite population S of size 40 to be the sample space. Convince yourself that from the point of view of probability distributions, the nature of the elements of S is irrelevant. Denote the element in the population corresponding to the i-th observation by W(i), i, = 1, …, 40. If one samples an element at random, W becomes a random variable. Which is its probability distribution?
b. Let X and T be two variables with X(i) = xi and T(i) =ti. Construct an example with the categorical variables W = student, X = school, and T = school district. Check that the relative frequency distributions of X and T are usually not constant. The experiment will then be to choose one student at random.
c. Compute the probability functions of X and T.

d. Why is it possible to obtain the probability function of T, if we are only given the probability function of X. *Hint:* Use additivity.
e. In this population model, all probabilities must take the form r/N. Note however, that when drawing a point at random in the unit square, the probability that it falls inside the inscribed circle is π/4. Can you represent this probability model exactly by an urn model? What happens if the population size N and the sample size n are very large and n/N is small.

DISCUSSION

This paper has discussed some interrelationships between several concepts in statistics, mathematics, and probability. In particular, in *Statistics*: population, sample, frequency distribution, dotplot, order statistics; in *Mathematics*: function, domain, codomain, extension, restriction, rational numbers, real numbers, continuity, countability, order relationship, permutations, and combinations with repetitions; in *Probability*: sample space, probability function, random variables, discrete and continuous, probability distribution. Our main objective has been to make some concepts in descriptive statistics more interesting and relevant for mathematically oriented people. One key idea is that a statistical variable provide good examples of the abstract notion of function. On the other hand, representing a variable by a function helps statistics students to see some subtle problems in variable type. Finally, we have argued that the teaching of random variables and their distributions can benefit a great deal from a previous study of statistical variables, which arise in descriptive statistics.

REFERENCES

del Pino, G. (2014). Invariance and descriptive statistics. Proceedings of 9th ICOTS. Makar, K. ed. Flagstaff, Arizona. July 13-18.

Devore, J. L. (2011). *Probability and Statistics for Engineering and the Sciences.* 8 ed. Cencage Learning.

Jacobs, H. R. (1979). Elementary Algebra. W. H. Freeman.

Montgomery, D. C. & Runger, G. C. (2010). *Applied Statistics and Probability for Engineers*. New York, Wiley.

Moore, D. S. (2010). *The basic practice of statistics*. New York: Freeman.

Usiskin, Z. (2014). On the relationship between statistics and other subjects in the K-12 curriculum. Proceedings of the 9th lnternational Conference on Teaching Statístics. Makar, K. ed. Flagstaff, Arizona. July 13-18.

Utts, J. M. & Heckard, R. F. (2007). *Mind on Statistics*. Cencage Learning

Walpole, R. E., Myers, R. H. & Myers, S. L. (2011). *Probability and Statistics for Engineers and Scientists* (9th Edition). Pearson.