THE RANDOM VARIABLE CONCEPT IN INTRODUCTORY STATISTICS

Theodore K. Miller, Indiana University, USA

From a general point of view, the concept of random variable is clearly fundamental to the field of statistics. Yet most introductory texts make almost no use of this idea, and those that do often use it in a way which creates problems in more advanced work. Moore (1995) relegates discussion of random variables to an optional section (Section 4.2), so that many students will never even see the phrase. Those that do see it, however, are likely to be quite confused because Moore uses the term in two very different ways. The first of these indicates that a statistic is a random variable, and this certainly makes sense in light of the definition of random variable used by Moore (1995, 250): "...a variable whose value is a numerical outcome of a random phenomenon". However, in the examples presented in Section 4.2, Moore also suggests that data values can be associated with the random variable idea. The first example focuses on a distribution of grades in a class, with a grade represented by a numeric value, while the second deals with the distribution of the household size variable. Moore makes the connection to the random variable idea by observing that if students are selected randomly one after the other (or households in the second example), the grade observed for a student (or the size of the household) will change.

The purpose of this paper is to argue that the association of data values with the idea of a random variable should be avoided. The first reason for doing so is that a data value does not clearly fit the definition of a random variable. Using Moore's definition, for example, a data value such as the size of a household is not a numerical outcome of a random phenomenon. The result of the random phenomenon (i.e. random sampling) is the particular household selected. That household has a size value that is quite independent of its random selection. A household not selected also has a size. This is very different from the case of a statistic, where the value clearly depends on the random selection of cases from the population. The value of a statistic does not exist outside the context of a randomly selected sample. The second reason for not associating data with the idea of a random variable is that sooner or later you have to recant. If you teach regression from an econometric point of view, for example, an important focus is on statistical properties of the OLS estimator, which is unbiased under certain conditions. One of these is the statistical independence of the error and the explanatory variables. When is this independence reasonably assumed to exist? The most common example is when the explanatory variables are viewed as fixed, and this is the condition assumed in the standard textbooks when the regression model is introduced. A good example is found in Green (1990, 147). The view of a data value as the value of a random variable in some contexts and of a fixed variable in others certainly has the potential to confuse students. The view presented in this paper is that a data value should be viewed as the value of a fixed variable unless there is a compelling reason to do otherwise. A value of the dependent variable in regression is the value of a random variable, for example, because it is defined in terms of a random error.

REFERENCES

Green, W. H. (1990). Econometric Analysis. MacMillian: New York.
Moore, D. S. (1995). The Basic Practice of Statistics. Freeman: New York.