Do we need only summary statistics or full datasets?

In a statistics course based on using theoretical-distribution methods, usually the first chapter describes how to find summary statistics and graphs, and then, in the remaining part of the course, we use only those summary statistics. But, when using simulation methods, we need the full dataset to do the simulations. The following chart is provided to help you think about these differences.

Type of data / parameters and (Usual theoretical dist'n method)	What is needed for **formula-based methods**	What is needed **for simulation methods**	**Is different information needed?**
Single Proportion (Normal dist'n methods)	Sample size and sample proportion	The entire set of 0's and 1's, but, of course, that can be recreated at any time from just the number of trials and the number of "successes."	Same information needed for both methods.
Single mean (t dist'n methods)	Sample size, sample mean, and sample standard deviation	Entire dataset	Different information needed for the two methods
Difference of two proportions from two independent groups* (Normal dist'n methods)	The fact that they are independent and then also, for EACH sample, the same as for a single proportion.	The fact that they are independent and then also, for EACH sample, the same as for a single proportion.	Same information needed for the two methods
Difference of two means from different independent groups (t dist'n methods)	The fact that they are inde pendent, and for EACH sample, the same as for a single mean	The entire dataset, with each number associated with the "group" it is in. Realistically, that means the dataset must be presented at a "stacked" dataset.**	Different information needed for the two methods
Two means, from "matched pairs" data (t dist'n methods)	Find difference for each pair.*** Use summary statistics on differences.	Use entire set of pairs.	Different information for the two methods.
Differences between multiple means from different independent groups (ANOVA for means)	Same as difference between two means	Same as difference between two means	Different information needed for the two methods
Multiple proportions, "dependent"* in some way (The two usual chi-squared tests)	Same as difference between two proportions.	Same as difference between two proportions.	Same information needed for the two methods.

* Proportions from "independent" groups have denominators that are counts of different groups. Ref. Lock Sec. 6.3D first paragraph

** A "stacked" dataset has each individual's information in one row and the variables as columns. (There may be one or more ID columns, which aren't considered variables, but just identifiers of the individual.) Almost all datasets in the Lock website are in this form. Ref: Lock, Sec. 1.1

*** To analyze "matched pairs" data, we can't use a traditional "stacked" dataset, because it doesn't give a straightforward way designate the pairing, which is needed to compute the necessary differences. In an elementary statistics course, we simply give the data in a table that makes it easy to compute the difference for each pair and collect those differences as values on one variable. Ref. Lock Sec. 6.5