**R basics**

M. Veeraraghavan

Aug. 24, 2014

**Data vector**: an array of single elements. Can ask for length of a data vector, not dim. A data vector can consist of numeric, character or logical values (http://www.statmethods.net/input/datatypes.html).

**Matrix**: rectangular grid of data - all of the same type. Can be a character matrix or a data matrix (with only numeric data). An **array** is a matrix with > 2 dimensions.

**Data frame**: rectangular grid of data - each column must contain the same type of data, as they are data vectors. But different columns can be of different types. For example, the first column can numeric while the Second column can be character.

**List**: most general form (flexible) of data storage. Collection of components. Each component can be a data vector or data frame or list of any length. A "data vector" is a "list" in which the top-level components are given by equal-length data vectors.

**Factor**: R uses factors to store categorical data. Factor is a different storage type from data vector. When a variable is declared a factor, its components are not numeric. They are categorical data, such as Yes or No. Levels are printed out at the bottom whenever a factor variable is printed. While a data vector can store factors, the function is.vector(x) will return FALSE if x is a data vector of factors.

**Types of data**: categorical, discrete numeric and continuous numeric. Besides numeric, data can be character or logical.

**Brackets**: Remember that R uses parentheses () for functions and square brackets [] for data objects. Lists use double square brackets [[]]

**Contingency tables**: In statistics, a contingency table (also referred to as **cross tabulation** or cross tab) is often used to record and analyze the relation between two or more categorical variables. It displays the (multivariate) frequency distribution of the variables in a matrix format. [From wikipedia]; “Bivariate categorical data is often presented in the form of a (two-way) contingency table.” The `table()` command takes unsummarized categorical data as in prev and grade vectors (which store factors, i.e., levels), and gives the frequencies of occurrences of the cross tabulation as numeric values.

**Functions**: First question ask how many mandatory parameters are needed as arguments for the function (later learn the options). Then ask what type of parameters are allowed - data vectors, matrices, categorical data (factor), numeric etc. For example, `boxplot(numeric data vector); barplot(table(categorical data vector)), dim(matrix), length(data vector), hist(numeric data vector)`

**Distribution**: skewed right (positive skewed) means long tail on the right, probability mass on the left, mean > median (e.g., incomes); opposite is true for distributions skewed left (e.g., final exam scores).