G DATA STRUCTURE AND STATISTICAL IMPLICATIONS

1 QUALITATIVE VARIABLES WITH CATEGORICAL OR DISCRETE DATA

`help tabulate`

There are two general types of variables: quantitative and qualitative. In a broad sense, quantitative data varies in magnitude (*e.g.* net income, price, interest rates) where qualitative data varies in kind (*e.g.* country, industry, codes, accounting standards). Examples of qualitative variables include the following:

(a) Categorical string data, such as:

  (i) Nominal string variables (*e.g.* `name` takes the values of firms' names, `j` takes the values of jurisdictional domicile for each firm, `curr` takes the values *DKK*, *EUR*, *GBP*, *SEK*, *USD*, and `i` assigns database codes to each firm).

  (ii) Ordinal variables (*e.g.* a string variable that distinguishes whether `ni` is *loss*, *zero*, or *profit*).

  (iii) Preference and multiple response variables resulting from surveys, etc.

(b) Discrete numerical data, such as:

  (i) Groupings (*e.g.* `g1`=0,1,2,…,9 and `g4`=100,…,9999).

  (ii) Counts (*e.g.* how many years a firm appears in the dataset, or how many losses a firm suffers in a period of 5 years).

  (iii) Binary variables (*e.g.* a variable that takes the value of 1 if `ni`≤0 and the value of 0 if `ni`>0).

  (iv) Quantile variables (*e.g.* percentiles, quartiles, bins for constructing histograms) etc.

(c) The identifier of time, such as `t`=2000,2001,…,2004. This is special type of qualitative variable that is necessary for identifying time steps when performing time series analysis or panel data analysis.

Some key characteristics of qualitative variables are:

(a) There are different levels of cross-section and each level may have fixed qualitative variables that remain invariant to time (*e.g.* the name of the firm and the code assigned to that firm by a database is fixed to the cross-sectional level of firm and does not change; currency is fixed to the cross-sectional level of jurisdiction and does not change in time etc).

(b) Some qualitative variables that are fixed to the level of cross-section may vary in time but very ifrequently, so that when they change status they indicate a kind of structural break (*e.g.* industry classification may change if the firm changes area of operations by merging or when acquired by another firm; domicile may change when a firm delists from one stock exchange and moves to another foreign stock exchange; even the company name may change).

(c) Other qualitative variables are not fixed to any cross-section and generally vary in time, such as dummy variables and quantile variables.

Qualitative variables are most useful when their categories or discrete data is tabulated. Stata offers numerous options for creating simple and complex tables, where tabulation can be expressed in nominal values, frequencies, accumulated frequencies or proportions. Managing continuous variables is discussed in the next section.