data types in statistics

Canonical name	DataTypesInStatistics
Date of creation	2013-03-22 14:44:27
Last modified on	2013-03-22 14:44:27
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	15
Author	CWoo (3771)
Entry type	Topic
Classification	msc 62-07
Defines	response variable
Defines	explanatory variable
Defines	continuous variable
Defines	discrete variable
Defines	categorical variable
Defines	nominal variable
Defines	ordinal variable
Defines	predictor
Defines	control variable
Defines	observation
Defines	qualitative variable
Defines	quantitative variable
Defines	dichotomous
Defines	polychotomous

Data drives statistics. In traditional statistical analysis, data can usually be visualized by a matrix. Each column in the matrix represents a data variable (slightly different from the mathematical definition of a variable), and each row respesents an observation or outcome, in which case only one data variable is involved, or a vector of observations or outcomes where several data variables are involved.

The types of data that are being distinguished have to do with the data variables. Before going into the details, let's begin with an example as a setting. A statistical analysis is conducted based on an observational study of autombile insurance data during a particular calendar year Yr. A matrix of data is formed with the following data variables being observed:

whether a policy has been involved in an accident during Yr,

NumAcc number of accidents have a policy been involved in an accident during Yr,

Cost the total amount of money a policy cost the insurance company during Yr,

Gen gender of driver,

Mar marital status of driver,

Age age of driver,

Hist number of accidents a driver had prior to year Yr,

DrvZIP zip code location where the driver lives,

AccZIP zip code location where the accident happened,

AccSt a numerical code corresponding to the state or province where the accident took place (for example, 0=Alabama, 1=Alaska, etc..., 50=Wyoming),

Inj the extent of the injury sustained during an accident,

VehType the type of vehicle in the policy, and finally,

VehWgt the weight of the vehicle in the policy.

Now, we are ready to breakdown the data variables. First, the data variables can be broken down in terms of their uses:

1. *response variable* or *predicted variable*. From the above example, , **NumAcc**, **Cost** can all be response variables. These are variables that we are trying to study, and predict.

2. *explanatory variable* or *predictor variable* or *control variable*. In the example above, given the response variable is , the explanatory variables can be any of the other variables except **NumAcc**, **Cost**, and **Inj**. Although possibly highly correlated with , **NumAcc**, **Cost**, and **Inj** do not "explain" why an accident occurs. In particular, **Inj** is only valid when there was an accident.

Usually, the response variable(s) $\boldsymbol{y}$ and the explanatory variable(s) $\boldsymbol{x}$ can be related functionally as

$$\boldsymbol{y} = f(\boldsymbol{x}).$$

A breakdown of data variables in terms of the natures of the variables is as follows:

1. *categorical variable* or *discrete variable*. These are data variables whose ranges are countable, often finite. Any value of a categorical variable is called a *level*, or a *category*. For example, is a categorical variable whose values are "Yes" (to mean that at least an accident occurred during year Yr) and "No" (to mean otherwise). A categorical variable whose number of values is two is often called a *binary variable* or a *dichotomous variable*. A categorical variable that has more than two values is called a *multinomial variable* or a *polychotomous variable*. **DrvZip**, **Inj** (no injury, light, medium, serious injuries, or death), **VehType** (family sedan, sports coupe, etc...) and **NumAcc** are examples of a multinomial variable.

2. *continuous variable*. Any data variable that is not a categorical variable is a continuous variable. **Age** and **VehWgt** are both examples of continuous variables. In real situations, these continuous variables usually lie within a certain bounded interval or ball (in higher dimensions). For example, it is safe to say that the range of the variable **Age** is $[\ 0, 140\ ]$.

In many statistical modeling situations, it is often convenient, sometimes even desirable to change continuous variables to categorical ones, and vice versa. Discretization is a way to turn a continuous variable into a categorical one. For example, the continuous variable **Age** can be turned into a dichotomous variable by the grouping: "Young" = Age $\in [0, 25]$ and "Not Young" = Age $\in (25, 140]$. Another possible grouping rule may be "Young" = Age $\in [0, 25]$, "Mature" = Age $\in (25, 55]$ Age and "Old" = Age $\in (55, 140]$.

Conversely, to turn a categorical variable into a continuous one, either the method of extension or transformation, or both, are used. For example, **Hist**, the number of prior accidents is a discrete variable taking on non-negative integer values, can be extended to a continuous variable taking on all non-negative real values to suit a certain modeling function f, even though non-integral values do not make sense and are not used in actual predictions. **AccZIP** can be transformed into a two-dimensional real-valued vector (longitude, latitude), since each (U.S.) zip code corresponds to an area with a unique centroid whose coordinate is measured in longitude and latitude.

Next, data variables can be grouped as whether they are:

1. *quantitative*. All variables such as **Age**, **NumAcc**, **Hist**, and **VehWgt** are quantitative variables since they take on numerical values. Variable **AccSt** is not a quantitative variable even though it is numeric in nature, since its values have no intrinsic numerical meanings. Another possible non-quantitative variable may be **DrvZIP**.

2. *qualitative*. Variables like **Gen**, **Mar**, **Inj**, as well as **AccSt** and **DrvZIP** are all qualitative variables.

Finally, data variables can be classified in terms of whether they can be ordered or not:

1. *nominal* variables have no intrinsic ordering structure. **Gen** and **Mar** are such examples, as are **AccSt**, **DrvZIP** and **VehType**.

2. The meaning of *ordinal* variables is self-explanatory. Usually, numerical variables are ordinal, except when they are multi-dimensional or vectorial. **AccZIP**, when transformed into longitude,latitude, is not ordinal. However, fixing any one of the two coordinates turns the other coordinate into an ordinal variable. An example of a non-numerical ordinal variable is **Inj**. Since the levels of **Inj** can be ranked by their severity, from "no injury" to "death", it is ordinal.

The data variables in the above example is summarized in the following table:

data variable	use	continuity	numerality	ordinality
	response	categorical	quantitative	nominal
NumAcc	response	categorical	quantitative	ordinal
Cost	response	continuous	quantitative	ordinal
Gen	explanatory	categorical	qualitative	nominal
Mar	explanatory	categorical	qualitative	nominal
Age	explanatory	continuous	quantitative	ordinal
Hist	explanatory	categorical	quantitative	ordinal
DrvZIP	explanatory	categorical	qualitative	nominal
AccZIP	explanatory	categorical	qualitative	nominal
AccSt	explanatory	categorical	qualitative	nominal
Inj	explanatory	categorical	qualitative	ordinal
VehType	explanatory	categorical	qualitative	nominal
VehWgt	explanatory	continuous	quantitative	ordinal