Common Data Types

Venki Uddameri, Ph.D., P.E.

Department of Civil, Environmental and Construction Engineering
Texas Tech University, Lubbock, TX 79409-1023
email: venki.uddameri@ttu.edu

September 8, 2019

Abstract

This short note discusses the various data types that civil engineers encounter. A basic understanding of the various data types is essential to identify appropriate models and analysis methods. Therefore, it is critical that you have a comprehensive understanding of the types of data.

1 Main Flavors of Data

At a very high level two flavors of data exist - 1) **quantitative** and **qualitative**.

1.1 Quantitative Data

Quantitative data deals with numbers and refers to variables that can be measured objectively. The load on a building, the width of a channel, the temperature of lake, and the bearing capacity of a soil are some examples of quantitative data.

The meaning of the quantitative data is clear and is interpreted the same way by all. However, while quantitative data can be measured objectively there can be some uncertainty associated with them due to measurement limitations (i.e, measurement errors).

1.2 Qualitative Data

Qualitative data refers descriptors that are subjective in nature. Linguistic descriptors such as *small* loads, *wide* channel, *sharp* turn and *extreme drought* are some examples of qualitative data. The subjective arises because what is small load for a large structure can be a heavy load for a much smaller structure.

The meaning of the qualitative data largely depends upon the situation and its interpretation can vary widely from one person to the next. Qualitative variables sometimes can be made quantitative for analysis. For example, **flood** is a qualitative variable. However, we can define a threshold flowrate to determine the flooding and non-flooding states in a river section.

2 Data Scales

There are four basic measurement scales of data, namely - nominal, ordinal, interval and ratio scales.

2.1 Nominal Scale

The nominal scale is the simplest of all data scales. These data are qualitative and often used to label the data. For example, a traffic survey could ask a respondent's gender (i.e., whether they are 'Male' or 'Female'). The category 'Gender' has two labels (Male and Female) which can be be used to filter data for further analysis.

Nominal scale data have no order or hierarchy associated with them. We cannot rank nominal scale data in any order.

2.2 Ordinal Scale

The ordinal scale involves arranging data in an order. This ordering can be used to rank or compare two or more values. Ordinal scale also refers to qualitative data. For example we could classify droughts as 'mild', moderate' and 'severe'. Clearly severe droughts have far greater impacts than moderate droughts which in turn have a greater impact than mild droughts.

We can also assign numeric values (say, 1 for mild, 2 for moderate and 3 for severe) to rank drought events in a region. However, we cannot do any arithmetic on these data (mild + moderate $\neq$ severe). Similarly the difference between a mild and moderate drought is not the same as the difference between a moderate and severe drought.

Ordinal scales are often used to assess relative perceptions, choices and feedback.

3 Interval Scale

Interval scale data have an order and it is possible to calculate exact differences between the values. The interval scale is used with quantitative data. You can calculate summary measures (e.g., mean, variance) on interval scale data.

The temperature measured in Celsius is a classic example of interval scale data. 20 ^oC is twice the value of 10 ^oC (i.e., the difference is 10 ^oC). Similarly, the difference between 90 ^oC and 80 ^oC is also 10 ^oC. This is because the values are being measured on the same interval.

However, 20 ^oC is not twice as hot as 10 ^oC. This is evident when these temperatures are converted to the Fahrenheit scale (10 ^oC = 50 ^oF and 20 ^oC = 68 ^oF). The reason for this is that there is no absolute zero in either Celsius of Fahrenheit scales. Zero just another measurement.

4 Ratio Scale

Variables measured on the ratio scale have an order, there are measured on fixed intervals (differences between two values can be calculated) and have an absolute zero. In other words, Ratio Scale has all the properties of an interval scale and also has an absolute zero. Ratio scale is used for quantitative data.

Consider daily rainfall amounts. There is an absolute zero which corresponds to no rainfall. Other measurable amounts of rainfall are measured with respect to this absolute zero. 20' of rainfall is more than 10' of rainfall. Also, 20' of rainfall is twice the amount of 10' accumulation. Their ratio is equal to 2. In a similar vein, 3' of rainfall is twice the amount of 6' accumulation (again the ratio is 2). As the name suggests, variables measured on ratio scale can be divided to obtain ratios or the degree of proportionality between two observations.

5 Continuous and Discrete Data

Quantitative data can be further split into continuous and discrete data.

Discrete data are **integers** and refer to variables that measure whole or indivisible entities. The number of vehicles passing an intersection over a year is discrete data as this is a whole number (integer).

Continuous data on the other hand are not restricted to integer values but also can take decimal values. For example, The temperature of water in a lake can be -67.3^oF.

Sometimes a variable that is continuous can be discretely measured or made discrete. For example, if we measure (or round off) the temperature of a lake to the nearest degree then we would have temperature measurements that are only integers. While continuous variables can be treated as discrete due to measurement limitations or data approximations, it is important to remember that intrinsically they can take continuous values and the discretization was simply a matter of convenience or a measurement limitation.

Summary measures of discrete variables (e.g., mean, standard deviation) can sometime have decimal values. For example, if 3 cars pass an intersection in the first hour and 4 cars in the second then the average number of cars is 3.5. In such instances, the value must be rounded appropriately (say to 4 if we are using it to estimate design loads on the pavement) as the decimal number is not physically realistic.

6 Continuous on a real line

Some variables in civil engineering are continuous on a real line (axis). For example, soil temperature measured in Celsius or Fahrenheit can assume both positive and negative values in cold regions.

Similarly, the groundwater table measured with respect to mean sea level (MSL) can be either positive (i.e., below MSL) or negative (above MSL). A zero value simply suggests that the water table coincides with MSL which is just as a valid state as water table being above or below MSL.

For variables that are continuous on a real line (axis), zero values have no special meaning and represent yet another value that the variable can take. In other words, they are often measured on an interval scale.

7 Positive Continuous Variables

As discussed above, the magnitude of a force is always positive. As this magnitude can take decimal values it is a positive continuous variable. Many variables of interest to civil engineers are positive continuous. Flowrates, pollutant concentrations, the magnitude of shear stresses are all positive continuous.

Some variables that are continuous on a real line can also be made positive continuous by shifting the datum. For example, there are no negative temperatures on Kelvin scale. Temperatures measured using Celsius scale can be made all positive by converting them into Kelvin scale. In a similar vein, the groundwater table measurements at a well can be made all positive by measuring values from the top of the well casing (assuming positive downwards) instead of the mean seal level.

8 Mixture Datasets

The positive continuous scale starts at zero and in theory extends all the way to infinity. For some positive continuous variables zero is yet another value if they are measured on the interval scale. On the other hand if they are measured on a ratio scale the value of zero is absolute and can assume special importance.

For example, zero rainfall implies no rainfall. Therefore, a time-series data of daily rainfall in a year can have several values of zero (no rainfall) and some days with positive (non-zero rainfall). Such a dataset can be viewed as a mixture - At a high level the data is discrete where each day belongs to either (rainfall=Yes) or (rainfall=No) state. Non-zero, positive continuous values exist for (Rainfall=Yes) state (see Figure 1).

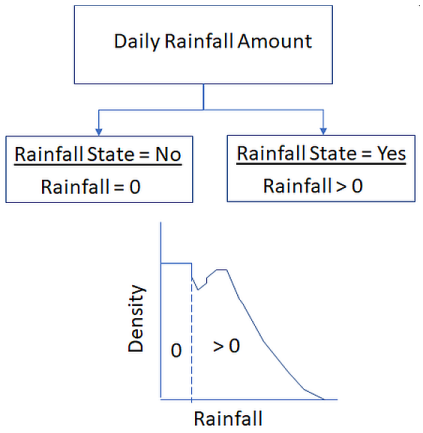

Figure 1: Illustration of Mixture Data

Similar mixtures can also be seen with discrete data. For example, if one were to count number of vehicles passing an intersection every hour of a day, there could be hours with no vehicles passing the intersection (traffic = No) and hours where some vehicles passed the intersection (traffic = Yes). A discrete dataset with non-zero values would correspond to the (traffic = Yes) state.

It is therefore important to ascertain upfront whether the value of zero is a valid measurement for the questions we seek to address and how zero values

will be handled during the analysis. Inclusion of a large number of zeros (zero-inflated datasets) will lower the mean and will affect other summary measures, on the other hand, exclusion of zero when it is a valid system response will lead to selection bias. The choices on how to handle zeros will depend upon the questions one seek to answer. It is important that the decision is made explicit and justified.

9 Truncated Data

Truncated data arise when values beyond a certain boundary are either not collected or removed prior to analysis. For example, the exclusion of zero values from a mixture data is one form of truncation.

Weigh-in-Motion (WIM)devices can measure axle weights and gross vehicle weights of trucks while in motion. WIM data is used to estimate loads on the bridge and also select vehicles for static inspection. Trucks weighing below a certain weight are allowed to pass through or even when their weights are collected, the data is discarded when ascertaining design loads. There is a threshold load boundary that is used to truncate the data.

Flooding risks are often ascertained using peak annual maximum flow data. While flowrates are measured at a much higher resolution (typically every 1 - 15 minutes), much of these data are discarded when assessing flooding risks. There is a threshold boundary that separates the maximum observed flow (state= max. flow) from the rest of the dataset (state = ¡ max. flow). However, the boundary in this case is not static and varies annually. For example, The maximum flow obtained in a dry year could be lower than say the tenth highest flow observed in a wet year.

Truncation of data is also common when conducting or analyzing surveys . Certain surveys may target a specific age-group (teenage drivers) while excluding others. Survey's focused on environmental justice issues tend to focus on low-income neighborhoods in industrial areas. These data subsets are often extracted from larger "quality of life" surveys that include respondents from different socioeconomic strata.

Truncation of data will lead to selection bias. It is therefore important that any truncation made explicitly during data analysis or implicitly during data collection be well documented. Any inferences drawn from the truncated data will be biased towards the sub-population that was sampled and will not be representative of the entire population. Therefore, caution must be exercised while interpreting inferences and insights from truncated data.

10 Censored Data

Censoring occurs when certain values in the dataset are only partially known. For example, a survey instrument measuring driving habits may classify the age group as ≤ 19 years. In this case, we know whether the respondent is below 20 years of age but we don't know the exact age of each respondent.

Censoring also arises due to instrument measurement limitations. Many instruments produce a small response even when there is on stimulus on the system. This response is referred to as the noise. The **detection limit** (DL) of an instrument is the lowest possible measurement that an instrument can make. This corresponds to a signal produced by the instrument that is significantly greater than the noise.

When an instrument measures a value below its detection limit we cannot be fully confident of the value as the signal is affected by the noise. It is common to report such values as $\leq$ DL.

Let us say, a flowmeter can reliably measure flows that are at least 2 cfs. A censored dataset would arise if some of the measurements are below 2 cfs. A censored dataset would like - [2.4, 3.2, $\leq$ 2.0, 2.1,$\leq$ 2.0,2.3,...]. Again, censored data is a form of mixture data where some numbers are known with greater certainty and reported as numbers and others are known with less certainty and reported as inequalities. **Left censoring** refers to the situation where values below a threshold are not known with certainty and represented as inequalities.

Many instruments are designed to work correctly to some specified upper limit. For example, a weighing scale may be designed to measure a maximum mass of 1000 kg. If we were to place a mass of 1200 kg on such a machine, we could ascertain that the mass is $\geq$ 1000 kg (instrument upper limit) but would not get a reading of 1200 kg. The dial on an analog scale would go past the maximum value of 1000 kg or a digital readout would throw an error message stating the mass is over the instrument maximum limit. Situations where values above a threshold are uncertain are referred to as **right censored data**. As another example, a traffic survey might categorize people over $\geq$ 70 years as elderly. Respondents who select these category are at least 75 years of age but we would not know how old they actually are.

It is not hard to conceive of situations where a dataset can be censored on both ends (i.e., be left and right censored). Such datasets are referred to as interval censored data

Figure 2 depicts the left and right censoring idea. The censoring occurs at values equal to 30 and 50. While the instrument may read a value of 25 or 55, because of the censoring they should be reported as $\leq$ 30 and $\geq$ 50 respectively.

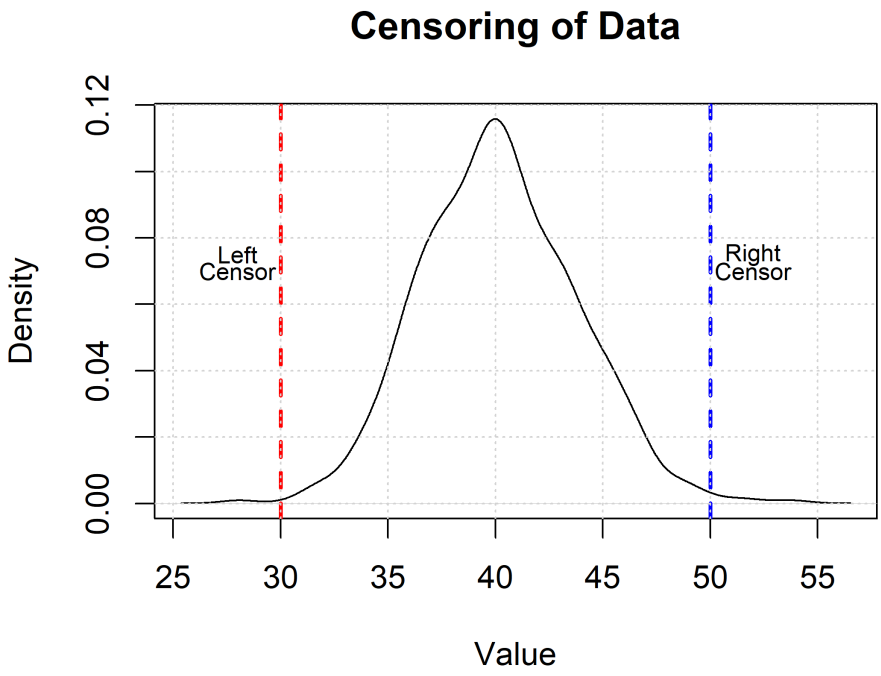

Figure 2: Left and Right Censored Data

11 Censoring versus Truncation

Censoring and truncation are distinct and should not be confused for one another. Both censoring and truncation define boundaries in the dataset. However, in censoring, we may be making (or have) measurements outside the boundaries. We don't want to throw away or exclude the data that are outside the censoring boundaries. These are valid measurements but we don't know their true values are.

On the other hand, in truncation, we are either not making measurements outside the boundaries or willingly excluding data that are collected outside the truncation boundaries. Both truncation and censoring are distinct from rounding where data are displayed to a specified number of significant digits to be consistent with the precision of the instrument and physical meaning of the variable.

12 Closing Remarks

Understanding basic data types and their measurement scales is an important first step of data analysis. The choice of the models and analysis techniques depend upon the characteristics of the data (data type and scale).