Rotteglada - Rosso, 2008

Chapter 1
Preliminary Data Analysis

All natural processes, as well as those devised by humans, are subject to variability. Civil engineers are aware, for example, that crushing strengths of concrete, soil pressures, strengths of welds, traffic flow, floods, and pollution loads in streams have wide variations. These may arise on account of natural changes in properties, differences in interactions between the ingredients of a material, environmental factors, or other causes. To cope with uncertainty, the engineer must first obtain and investigate a *sample* of data, such as a set of flow data or triaxial test results. The sample is used in applying statistics and probability at the descriptive stage. For inferential purposes, however, one needs to make decisions regarding the *population* from which the sample is drawn. By this we mean the total or aggregate, which, for most physical processes, is the virtually unlimited universe of all possible measurements. The main interest of the statistician is in the aggregation; extrapolation the individual items provide the hints, clues, and evidence.

A data set comprises a number of measurements of a phenomenon such as the failure load of a structural component. The quantities measured are termed *variables*, each of which may take any one of a specified set of values. Because of its inherent randomness and hence unpredictability, a phenomenon that an engineer or scientist usually encounters is referred to as a *random variable*, a name given to any quantity whose value depends on chance.[1] Random variables are usually denoted by capital letters. These are classified by the form that their values can possibly take (or are assumed to take). The pattern of variability is called a *distribution*. A *continuous* variable can have any value on a continuous scale between two limits, such as the volume of water flowing in a river per second or the amount of daily rainfall measured in some city. A *discrete* variable, on the contrary, can only assume countable isolated numbers like integers, such as the number of vehicles turning left at an intersection, or other distinct values.

e.g. INFERENCE

Having obtained a sample of data, the first step is its presentation. Consider, for example, the modulus of rupture data for a certain type of timber shown in Table E.1.1, in Appendix E. The initial problem facing the civil engineer is that such an array of data by itself does not give a clear idea of the underlying characteristics of the stress values in this natural type of construction material. To extract the salient features and the particular types of information one needs, one must summarize the data and present them in some readily comprehensible forms. There are several methods of presentation, organization, and reduction of data. Graphical methods constitute the first approach.

1.1 GRAPHICAL REPRESENTATION

If "a picture is worth a thousand words," then graphical techniques provide an excellent method to visualize the variability and other properties of a set of data. To the powerful interactive system of one's brain and eyes, graphical displays provide insight into the form

[1] The term will be formally defined in Section 3.1.