Chapter 2: Data

Data are numbers or characteristics with some specific meaning. (Data is the plural of datum.)

Example: Consider these data: 18, 22, 19, 21, 38.

- This is just a set of 5 numbers with little meaning.
- These could be the ages of five randomly sampled Stat 451 students.
- It could also be the speed (mph) of five different vehicles observed at the corner of Broadway and Higgins this morning.
- Data have no meaning without knowledge of their origin.

Data originate by measuring an attribute (e.g., age) or characteristic (gender) associated with an object, element or person.

Terminology

- Observational units are the things that are being observed, and from which measurements are obtained. (Example: Record the processing speed of each computer in a sample of campus computers; then, the computers are the observational units). When working with people, the observational units are sometimes called subjects, respondents, or participants. The question of who (is being observed), identifies the observational units.

- A variable is an attribute or characteristic of an observational unit. The values of a variable vary among the observational units. The question of what is being measured, asks what variables are being observed.

- Data are values taken on by a variable measured on a set of observational units.

- Two types of variables are:

  1. Categorical (also called qualitative or nominal) variables identify a category to which an observational belongs. Examples include gender and class standing. A categorical variable with only two possible categories is a binary variable.

  2. Quantitative variables take on numerical values. Numerical values can be ordered (e.g., smallest to largest) and arithmetic operations can be carried out sensibly. Counts, concentrations and percentages are examples of quantitative variables.
Examples:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter of an angiogram catheter</td>
<td>quantitative</td>
</tr>
<tr>
<td>Time spent on a particular web page</td>
<td>quantitative</td>
</tr>
<tr>
<td>Country of birth</td>
<td>categorical</td>
</tr>
<tr>
<td>Age of an individual</td>
<td>quantitative</td>
</tr>
<tr>
<td>Income level of an individual</td>
<td>categorical (ordinal)</td>
</tr>
<tr>
<td>Customer satisfaction level at the DMV</td>
<td>categorical (ordinal)</td>
</tr>
<tr>
<td>(1=poor, 2=OK, 3=good, 4=excellent)</td>
<td></td>
</tr>
</tbody>
</table>

- In the customer satisfaction example, a numerical score is used to represent the satisfaction level. Caution must be used in treating customer satisfaction as a quantitative variable.

- A variable which is categorical, but has a natural ordering (such as the satisfaction variable above) is *ordinal*. We will not consider methods for ordinal variables in this course.

- Methods of data summarization are visual or numerical in nature. *Visual* (graphical) methods are effective for rapid communication. *Numerical* summaries are effective obtaining deeper understanding of the population from which they were drawn.

- A *statistic* is a numerical summary of a data set. The *percentage* of graduate students and the *mean* length of time spent on a web page are statistics.

Understanding the origins and characteristics of a data set are essential for correct statistical analysis. It may be helpful to identify the who, what, where, when, why, and how of the data.

More formally, and throughout the course, the terms *observational units* and *variables* will be used. A *datum* is a measurement made on an *observational unit* and a variable is a characteristic or attribute that is being measured on the observational unit.

A data set consists of $p \geq 1$ measurements made on $n \geq 1$ observational units. If $p$ variables are measured on each unit, then it is expected that the data set will consist of $np$ measurements. The standard organization of data sets is $n$ rows (one row per observational unit) and $p$ columns (one column) per variable.
For each identified variable, identify the type as quantitative or categorical:

*Example 1:* A company concerned with low employee participation rate in its 401(K) plan, sampled 30 other companies with similar plans and asked for their 401(K) participation rates, and the percentage of salary matched by the employer.

Who?  
What?  
Where?  
When?  
Why?  
How?  
Observational Units?  
Variable(s) and Type(s)?

*Example 2:* In the mid-twentieth century, large volumes of hazardous materials disposed of by burying them near the Idaho National Laboratory. In a study to compare the vapor concentration of the carcinogen carbon tetrachloride before and after a pumping mechanism was installed, concentrations were collected at 8 depths down to 240 feet at each of 15 wells in a 1 mile-square area within the Radioactive Waste Management Complex of the Idaho National Laboratory. These concentrations were collected both before and after the pumping mechanism was installed.

Who?  
What?  
Where?  
When?  
Why?  
How?  
Observational units?  
Variable(s) and Type(s)?