3.3.4 Organizing With Descriptive Statistics

Descriptive statistics, about a collection or dataset, summarize it concisely and can identify the properties that might be most useful as organizing principles. The simplest statistical description of a collection is how big it is; how many resources or observations does it contain?

Descriptive statistics summarize a collection of resources or dataset with two types of measures:

- Measures of *central tendency*: Mean, median, and mode; which measure is appropriate depends on the level of measurement represented in the numbers being described (these measures and the concept of levels of measurements are defined in §3.3.3 Organizing Digital Resources (page 111)).
- Measures of *variability*: Range (the difference between the maximum and minimum values), and standard deviation (a measure of the spread of values around the mean).

Statistical descriptions can be created for any resource property, with the simplest being the number of resources that have the property or some particular value of it, such as the number of times a particular word occurs in a document or the number of copies a book has sold. Comparing summary statistics about a collection with the values for individual resources helps you understand how typical or representative that resource is. If you can compare your height of 6 feet, ½ inch with that of the average adult male, which is 5 feet, 10 inches, the difference is two and a half inches, but what does this mean? It is more informative to make this comparison using the standard deviation, which is three inches, because this tells you that 68% of adult men have heights between 5 feet, 7 inches and 6 feet, 1 inch. When measurements are normally distributed in the familiar bell-shaped curve around the mean, the standard deviation makes it easy to identify statistical outliers.

No matter how measurements are distributed, it can be useful to employ descriptive statistics to organize resources or observations into categories or quantiles that have the same number of them. Quartiles (4 categories), deciles (10), and percentiles (100) are commonly used partitions.

Alternatively, resources or observations can be organized by visualizing them in a histogram, which divides the range of values into units with equal intervals. Because values tend to vary around some central tendency, the intervals are unlikely to contain the same number of observations. Descriptive statistics and associated visualizations can suggest which properties make good organizing principles because they exhibit enough variation to distinguish resources in their most useful interactions. For example, it probably isn't useful to organize

books according to their weight because almost all books weigh between ½ and 2 pounds, unless you are in the business of shipping books and paying according to how much they weigh.

3.3.4.1 Exploratory Analysis to Understand Data

Many experts recommend that data analysts should undertake some exploratory analysis with descriptive statistics and simple information visualizations to understand their data before applying sophisticated computational techniques to the dataset. In particular, because the human visual system quickly perceives shapes and patterns, analyzing and graphing the values of data attributes and other resource descriptions can suggest which properties might be useful and comprehensible organizing principles. In addition, data visualization makes it easy to recognize values that are typical or that are outliers. Some of this analysis might form part of data quality assessment during resource selection, but if not done then, it should be done as part of the organizing process.

A dataset whose fields or attributes lack information about data types and units of measure has little use because the data lacks meaning. When some, but not all parts of the data are named or annotated, avoid over-interpreting these descriptions' meanings. (See §4.4 Naming Resources (page 203).)

We will do some exploratory analysis to understand what an example dataset contains and how we might use it. For our example, we consider a collection of a few hundred records from a healthcare study, whose first eight records and first five data fields in each record are shown in Figure 3.2a, Example Dataset.

Figure 3.2a. Example Dataset

ID	Sex	Temp	Age	Weight	...	...	...	...	...
1	1	97.6	32	135					
2	0	97.6	19	118					
3	0	97.6	23	128					
4	1	98.7	34	140					
5	1	98.5	52	162					
6	1	98.7	60	160					
7	0	98.3	36	148					
8	0	98.3	38	155					
...	...								
260	1	99.0	23	123					

The "ID" column contains numeric data, but every value is a different integer, and the values are contiguous. The field label "ID" suggests that this is the re-

source identifier for the participants in the healthcare study. Further examination of other tables will reveal that this is a key value that points into a different dataset containing the resource names.

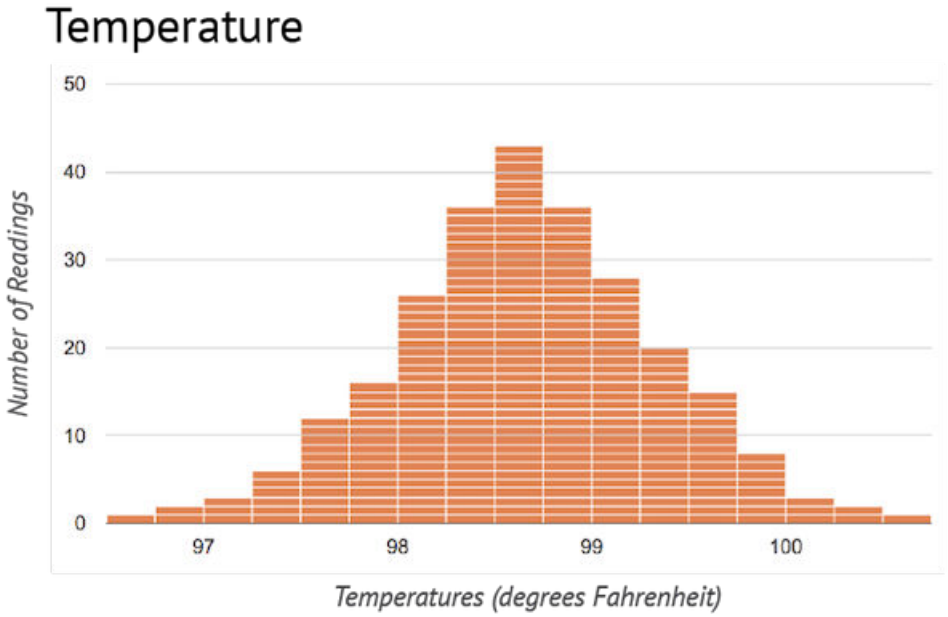
The "Sex" column is also numeric, but there are only two different values, 0 and 1, and in the complete dataset they are approximately equal in frequency. This attribute seems to be categorical or Boolean data. This makes sense for a "Sex" categorization, and it is likely to prove useful in understanding the dataset.

Histogram

A histogram is the simplest visualization of one-dimensional data. It is a bar graph that takes the full range of values, organizes them into a set of intervals of equal size on one axis, and then counts the number of values in each interval on the other axis.

The "Temp" column contains several hundred different numeric values in the complete dataset, ranging from 96.8 to 100.6, with a mean of 98.6. These values are sensible if the label "Temp" means the under-the-tongue body temperature in degrees Fahrenheit of the study participant when the other measures were obtained. This type of data is usefully viewed as a histogram to get a sense of the spread and shape, shown in Figure 3.2b, Temperature.

Figure 3.2b. Temperature



The data values of the "Temp" column follow the familiar normal or bell-shaped distribution, for which simple and useful descriptive statistics are the mean and the standard deviation. The mean (or average) is at the center of the distribution, and the standard deviation captures the width of the bell shape. In this dataset, the very narrow range of data values here suggests that this attribute is not useful as an organizing principle, since it does not distinguish the resources in any significant way. In a larger sample, however, there might be a few very low or very high temperatures, and it would be useful to investigate these "hypothermic" or "hyperthermic" outliers.

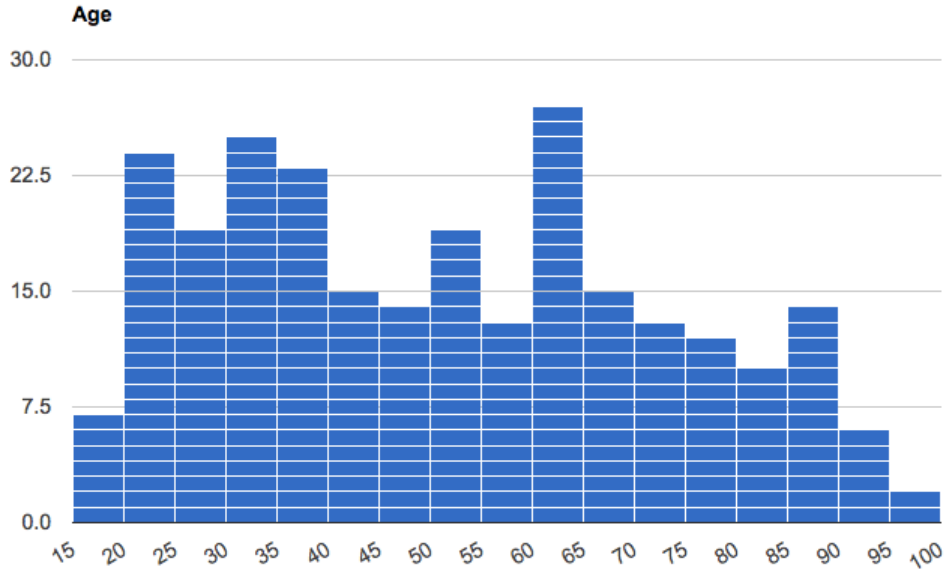
The data values of the "Age" column range from 18 to 97, and are spread broadly across the entire range; this is the age, in years, of the study participants. When a distribution is very broad and flat, or highly skewed with many values at one end or another, the mean value is less useful as a descriptive statistic. Instead of the mean, it is better to use the median or middle value as a summary of the data; the median value for "Age" in the complete dataset is 39.

Median versus Average

If ten people are in a bar, all of whom make $50,000 a year, when a movie star who made $25,000,000 this year walks in, the average income is now $2.3 million. The median income is still $50,000.

The End of Average tells the story of how the U.S. military designed aircraft cockpits beginning in 1926 on the basis of the average dimensions of a 1926 pilot. In 1950, researchers measured over four thousand pilots only to discover that no actual pilot had average values on all the measures, and recommended adjustable seats and controls in cockpit design.[85][DS]

Figure 3.2c. Age



The "Weight" column has about 220 different numeric values, from 82 to 300, and judging from this range we can infer that the weights are measured in pounds. The data follows an uneven distribution with peaks around 160 and 200, and a small peak at 300. This odd shape appears in the histogram of Figure 3.2d, Weight. The two peaks in this so-called multi-modal histogram suggest that this measure is mixing two different kinds of resources, and indeed it is because weights of men and women follow different distributions. It would thus be useful to use the categorical "Sex" data to separate these populations, and Figure 3.2e, Sex and Weight: Female shows how analyzing weight for women and men as different populations is much more informative as an organizing principle than combining them.

What about the odd peak in the distribution at 300? End of range anomalies like this generally reflect a limitation in the device or system that created the data. In this case, the weight scale must have an upper limit of 300 pounds, so the peak represents the people whose weight is 300 or greater.

Figure 3.2d. Weight

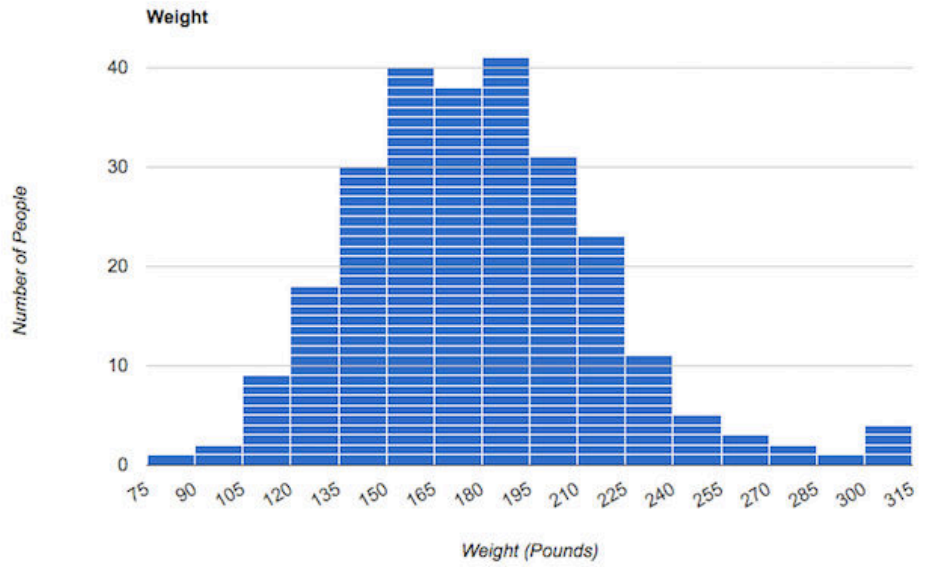


Figure 3.2e. Sex and Weight: Female

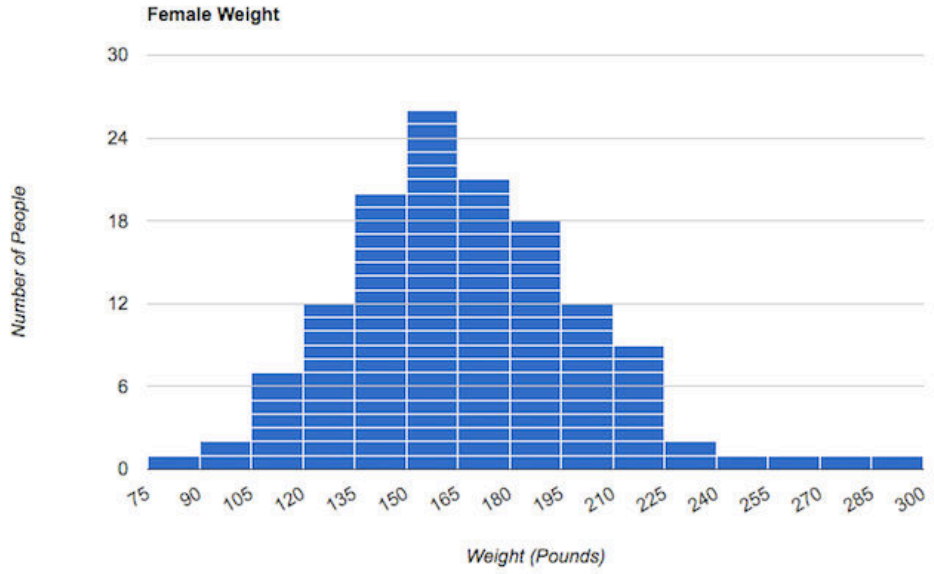
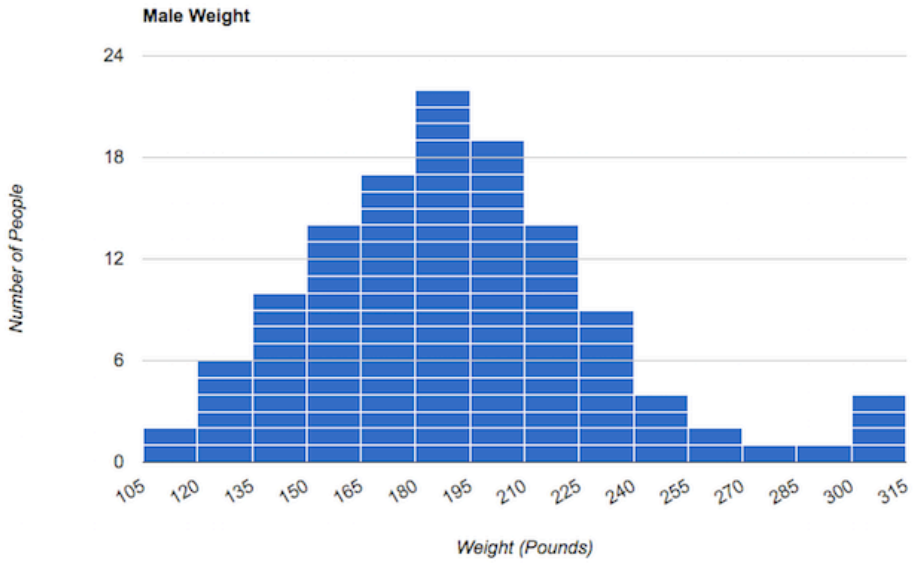


Figure 3.2f. Sex and Weight: Male



3.3.4.2 Detecting Errors and Fraud in Data

There are numerous techniques for evaluating individual data items or datasets to ensure that they have not been changed or corrupted during transmission, storage, or copying. These include parity bits, check digits, check sums, and cryptographic hash functions. They share the idea that a calculation will yield some particular value or match a stored result when the original data has not been changed. Another basic technique for detecting errors is to look for data values that are different or anomalous because they do not fall into expected ranges or categories.

More interesting challenges arise when the data might have been changed by intentional actions to commit fraud, launder money, or carry out some other crime. In these situations, the person tampering with data or creating fake data will try to make the data look normal or expected.

Forensic accountants and statisticians use many techniques for detecting possibly fraudulent data in these adversarial contexts. Some are quite simple:

- If expenses are reimbursed up to some maximum allowed value, look for data items with that exact value.
- When any value exceeding some threshold triggers more careful analysis, look for other data items just below that threshold.

- When invoices or claims are paid on receipt, and only a sample are subsequently audited, look for duplicate submissions.
- Calculate the ratio of the maximum to the minimum value for purchases in some category (such as the unit price paid for items from suppliers); items with large ratios might indicate fraud where the supplier "kicks back" some of the money to the purchaser.

Benford's Law, the observation that the leading digits in data sets are distributed in a non-uniform manner, is an effective technique for detecting fraudulent data because it is based on a counter-intuitive fact not known to most fraudsters, who often make up data to look random. You might think that the number 1 would occur 11% of the time as the first digit (since there are 9 possibilities), but for data sets whose values span several orders of magnitude, the number 1 is the first digit about 30% of the time, and 7, 8, and 9 occur around 5%.

Because of the very high transaction rate and the relatively small probability of fraud, credit card fraud is detected using machine learning algorithms. The classifier is trained with known good and bad transactions using properties like average amount, frequency, and location to develop a model of each cardholder's "data behavior" so that a transaction can quickly be assigned a probability that it is fraudulent. (More about this kind of computational classification in Chapter 7, *Categorization: Describing Resource Classes and Types*.)[86[DS]]

3.3.5 Organizing with Multiple Resource Properties

Multiple properties of the resources, the person organizing or intending to use them, and the social and technological environment in which they are being organized can collectively shape their organization. For example, the way you organize your home kitchen is influenced by the physical layout of counters, cabinets, and drawers; the dishes you cook most often; your skills as a cook, which may influence the number of cookbooks, specialized appliances and tools you own and how you use them; the sizes and shapes of the packages in the pantry and refrigerator; and even your height.

If multiple resource properties are considered in a fixed order, the resulting arrangement forms a *logical hierarchy*. The top level categories of resources are created based on the values of the property evaluated first, and then each category is further subdivided using other properties until each resource is classified in only a single category. Consider the hierarchical system of folders used by a professor to arrange the digital resources on his computer; the first level distinguishes personal documents from work-related documents; work is then subdivided into teaching and research, teaching is subdivided by year, and year divided by course.