2.1: Ways to Describe Data

If we are to consider how to describe data, then we need some data with which we can work. Ideally, we want data that is easy to gather and easy to understand. It also is helpful if you can gather similar data on your own so you can repeat what we cover here. A simple system that meets these criteria is to analyze the contents of bags of M&Ms. Although this system may seem trivial, keep in mind that reporting the percentage of yellow M&Ms in a bag is analogous to reporting the concentration of Cu^{2+} in a sample of an ore or water: both express the amount of an analyte present in a unit of its matrix.

At the beginning of this chapter we identified four contrasting ways to describe data: categorical vs. numerical, ordered vs. unordered, absolute reference vs. arbitrary reference, and discrete vs. continuous. To give meaning to these descriptive terms, let's consider the data in Table 2.1.1, which includes the year the bag was purchased and analyzed, the weight listed on the package, the type of M&Ms, the number of yellow M&Ms in the bag, the percentage of the M&Ms that were red, the total number of M&Ms in the bag and their corresponding ranks.

Table 2.1.1. Distribution of Yellow and Red M&Ms in Bags of M&Ms.

bag id	year	weight (oz)	type	number yellow	% red	total M&Ms	rank (for total)
a	2006	1.74	peanut	2	27.8	18	sixth
b	2006	1.74	peanut	3	4.35	23	fourth
c	2000	0.80	plain	1	22.7	22	fifth
d	2000	0.80	plain	5	20.8	24	third
e	1994	10.0	plain	56	23.0	331	second
f	1994	10.0	plain	63	21.9	333	first

The entries in Table 2.1.1 are organized by column and by row. The first row—sometimes called the header row—identifies the variables that make up the data. Each additional row is the record for one sample and each entry in a sample's record provides information about one of its variables; thus, the data in the table lists the result for each variable and for each sample.

Categorical vs. Numerical Data

Of the variables included in Table 2.1.1, some are categorical and some are numerical. A categorical variable provides qualitative information that we can use to describe the samples relative to each other, or that we can use to organize the samples into groups (or categories). For the data in Table 2.1.1, bag id, type, and rank are categorical variables.

A numerical variable provides quantitative information that we can use in a meaningful calculation; for example, we can use the number of yellow M&Ms and the total number of M&Ms to calculate a new variable that reports the percentage of M&Ms that are yellow. For the data in Table 2.1.1, year, weight (oz), number yellow, % red M&Ms, and total M&Ms are numerical variables.

We can also use a numerical variable to assign samples to groups. For example, we can divide the plain M&Ms in Table 2.1.1 into two groups based on the sample's weight. What makes a numerical variable more interesting, however, is that we can use it to make quantitative comparisons between samples; thus, we can report that there are $14.4\times$ as many plain M&Ms in a 10-oz. bag as there are in a 0.8-oz. bag.

$$\frac{333+331}{24+22}=\frac{664}{46}=14.4$$

Although we could classify year as a categorical variable—not an unreasonable choice as it could serve as a useful way to group samples—we list it here as a numerical variable because it can serve as a useful predictive variable in a regression analysis. On the other hand rank is not a numerical variable—even if we rewrite the ranks as numerals—as there are no meaningful calculations we can complete using this variable.

Nominal vs. Ordinal Data

Categorical variables are described as nominal or ordinal. A nominal categorical variable does not imply a particular order; an ordinal categorical variable, on the other hand, coveys a meaningful sense of order. For the categorical variables in Table 2.1.1, bag

id and type are nominal variables, and rank is an ordinal variable.

Ratio vs. Interval Data

A numerical variable is described as either ratio or interval depending on whether it has (ratio) or does not have (interval) an absolute reference. Although we can complete meaningful calculations using any numerical variable, the type of calculation we can perform depends on whether or not the variable's values have an absolute reference.

A numerical variable has an absolute reference if it has a meaningful zero—that is, a zero that means a measured quantity of none—against which we reference all other measurements of that variable. For the numerical variables in Table 2.1.1, weight (oz), number yellow, % red, and total M&Ms are ratio variables because each has a meaningful zero; year is an interval variable because its scale is referenced to an arbitrary point in time, 1 BCE, and not to the beginning of time.

For a ratio variable, we can make meaningful absolute and relative comparisons between two results, but only meaningful absolute comparisons for an interval variable. For example, consider sample e, which was collected in 1994 and has 331 M&Ms, and sample d, which was collected in 2000 and has 24 M&Ms. We can report a meaningful absolute comparison for both variables: sample e is six years older than sample d and sample e has 307 more M&Ms than sample d. We also can report a meaningful relative comparison for the total number of M&Ms—there are

$$\frac{331}{24} = 13.8\times$$

as many M&Ms in sample e as in sample d—but we cannot report a meaningful relative comparison for year because a sample collected in 2000 is not

$$\frac{2000}{1994} = 1.003\times$$

older than a sample collected in 1994.

Discrete vs. Continuous Data

Finally, the granularity of a numerical variable provides one more way to describe our data. For example, we can describe a numerical variable as discrete or continuous. A numerical variable is discrete if it can take on only specific values—typically, but not always, an integer value—between its limits; a continuous variable can take on any possible value within its limits. For the numerical data in Table 2.1.1, year, number yellow, and total M&Ms are discrete in that each is limited to integer values. The numerical variables weight (oz) and % red, on the other hand, are continuous variables. Note that weight is a continuous variable even if the device we use to measure weight yields discrete values.
