Statistical Thinking for the 21st Century

Chapter 3

Summarizing data

3.2.2 Cumulative distributions

Often we wish to summarize data that can have many possible values. When those values are quantitative, then one useful way to summarize them is via what we call a *cumulative* frequency representation: rather than asking how many observations take on a specific value, we ask how many have a value of *at least* some specific value.

Let's look at a variable called SleepHrsNight which records how many hours a participant reports sleeping on usual weekdays. Let's create a frequency table as we did before. Such a frequency table is presented in Table 3.2 (on the next page).

From looking at our frequency table (Table 3.2, on the next page), we can already begin to summarize the dataset just by looking at the absolute and relative frequency. For example, we can see that most people report sleeping between 6 and 8 hours.

Table 3.2: Frequency distribution for number of hours of sleep per night in the NHANES dataset

SleepHrsNight	AbsoluteFrequency	RelativeFrequency	Percentage
2	9	0.00	0.18
3	49	0.01	0.97
4	200	0.04	3.97
5	406	0.08	8.06
6	1172	0.23	23.28
7	1394	0.28	27.69
8	1405	0.28	27.90
9	271	0.05	5.38
10	97	0.02	1.93
11	15	0.00	0.30
12	17	0.00	0.34

What if we want to know how many people report sleeping 5 hours or less? To find this, we can compute a *cumulative distribution*. To compute the cumulative frequency for some value j, we add up the frequencies for all of the values up to and including j:

$$cumulative\ frequency_j = \sum_{i=1}^{j} absolute\ frequency_i$$

Table 3.3, below, shows our cumulative frequency (as well as absolute frequency, which we had computed before). A cumulative frequency is like a running total. For example, 664 participants reported sleeping 5 hours or less. 4635 participants reported sleeping 8 hours or less.

Table 3.3: Absolute and cumulative frequency distributions for SleepHrsNight variable

SleepHrsNight	AbsoluteFrequency	CumulativeFrequency
2	9	9
3	49	58
4	200	258
5	406	664
6	1172	1836
7	1394	3230
8	1405	4635
9	271	4906
10	97	5003
11	15	5018
12	17	5035

Below, in the left panel of Figure 3.3 we plot the data to see what these representations look like; the absolute frequency values are plotted in solid lines, and **the cumulative frequencies are plotted in dashed lines.**

We see that the cumulative frequency is monotonically increasing – that is, it can only go up or stay constant, but it can never decrease.

Again, we usually find the relative frequencies to be more useful than the absolute; those are plotted in the right panel of Figure 3.3.

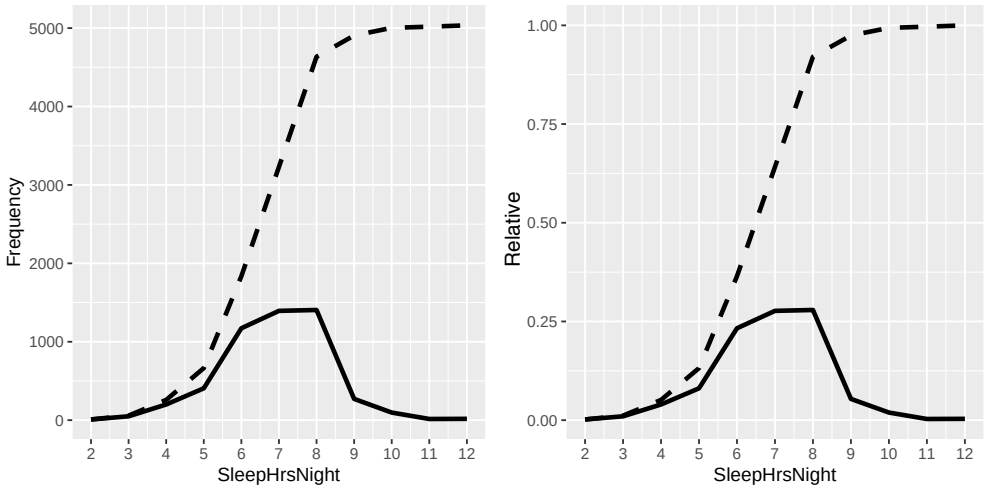

Figure 3.3: A plot of the frequency distribution (solid) and cumulative distribution (dashed) values for absolute (left) and relative (right) for the possible values of SleepHrsNight.