Creating a Histogram

This document shows you how to create a histogram from data. But it is also useful for understanding how to interpret a histogram as well since knowing how to interpret it requires knowing how it was made from the data.

Consider a sample of $n = 20$ observations of a *quantitative* variable.

5.2, 8.7, 11.6, 12.2, 12.2, 12.6, 12.8, 13.8, 15.4, 15.6, 15.6, 16.8, 17, 17.3, 17.5, 18.5, 21.1, 22.8, 23.6, 23.9

Note that the observations have been sorted so that they are in increasing order. This isn't strictly necessary but it makes the next steps easier. Next we need to select intervals. These intervals should cover all of the observations and should be of the same width. The location and width of the intervals are somewhat arbitrary, but the choice shoudl give a good representation of the distribution without having too many intervals. I have selected 10 intervals starting at 4 and ending at 24, each with a width of 2. These are shown in the following table.

Interval	Frequency	Relative Frequency	Density
4 to 6	1	0.05	0.025
6 to 8	0	0.00	0.000
8 to 10	1	0.05	0.025
10 to 12	1	0.05	0.025
12 to 14	5	0.25	0.125
14 to 16	3	0.15	0.075
16 to 18	4	0.20	0.100
18 to 20	1	0.05	0.025
20 to 22	1	0.05	0.025
22 to 24	3	0.15	0.075
	20	1.00	

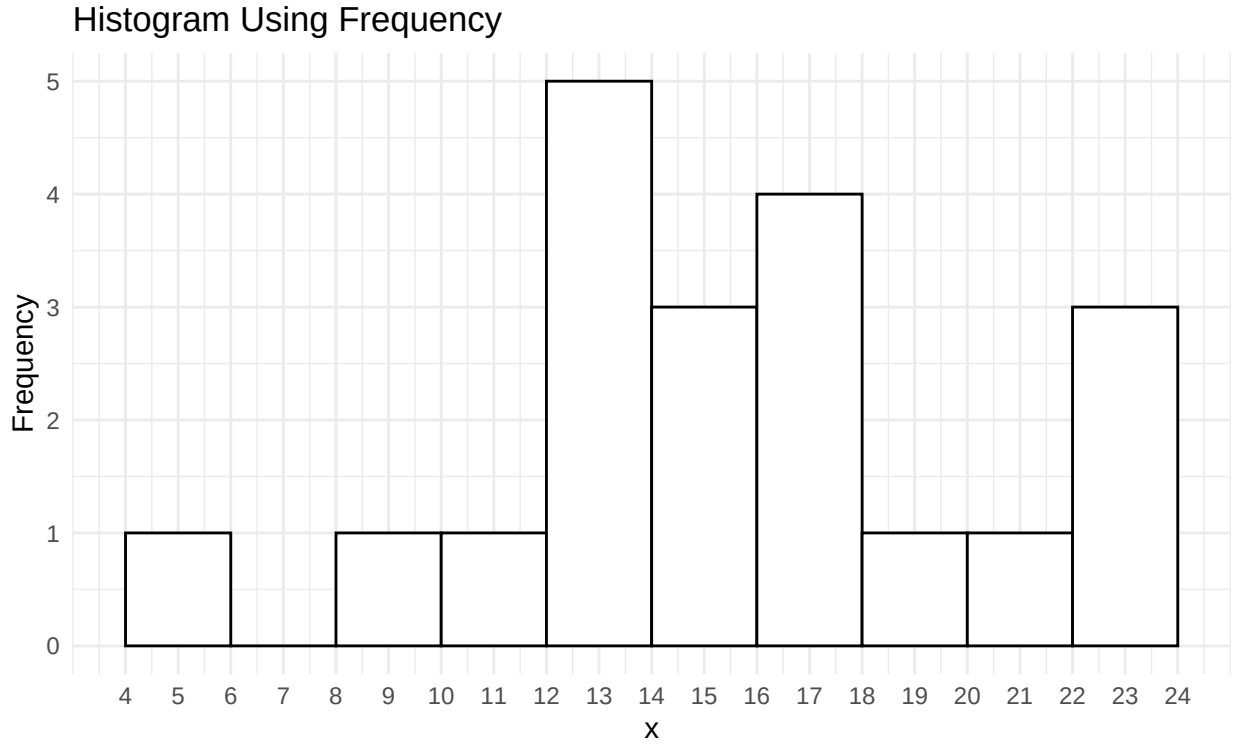
One minor detail to consider is that some observations might fall exactly on the break point between two intervals (e.g., an observation of 14 is on the break point between the interval 12 to 14 and the interval 14 to 16). We simply need a convention to resolve such observations. We will use the convention that such an observation will go in the *lower* interval (e.g., an observation of 14 will be counted as being in 12 to 14, not 14 to 16).

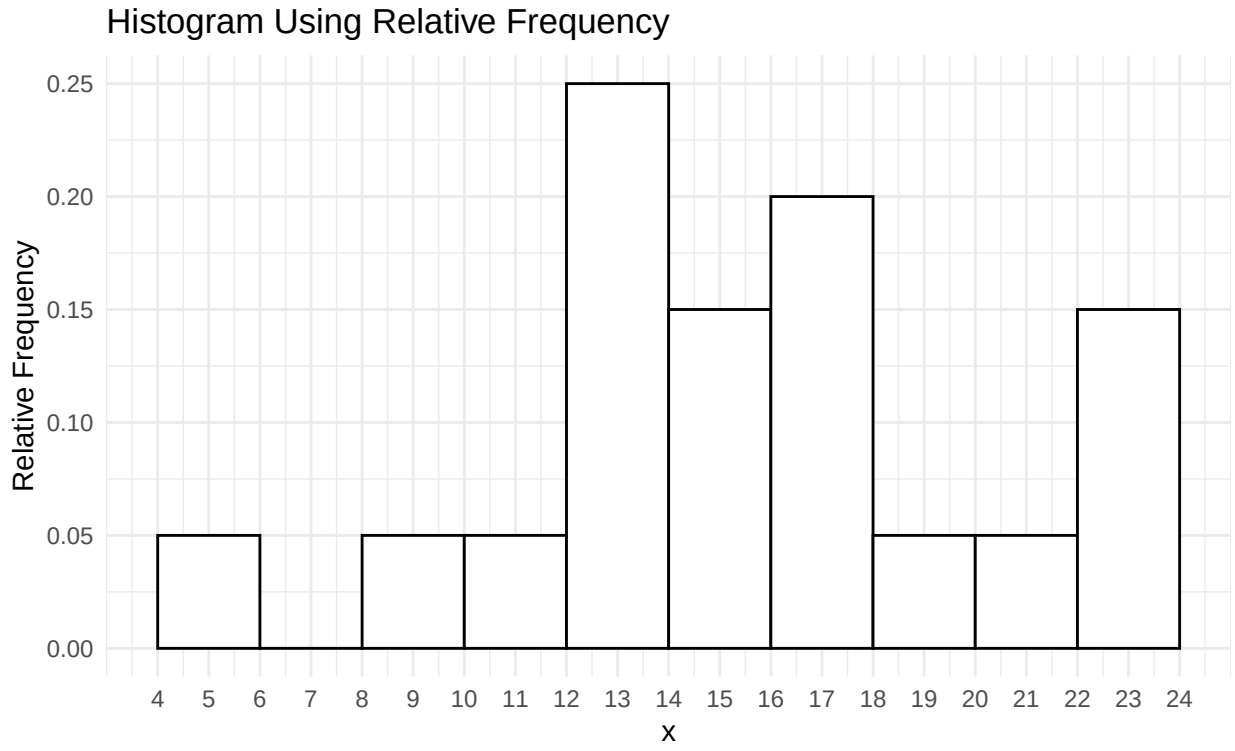
For each interval the table shows three ways to measure how often observations fall into each interval: *frequency*, *relative frequency*, and *density*.

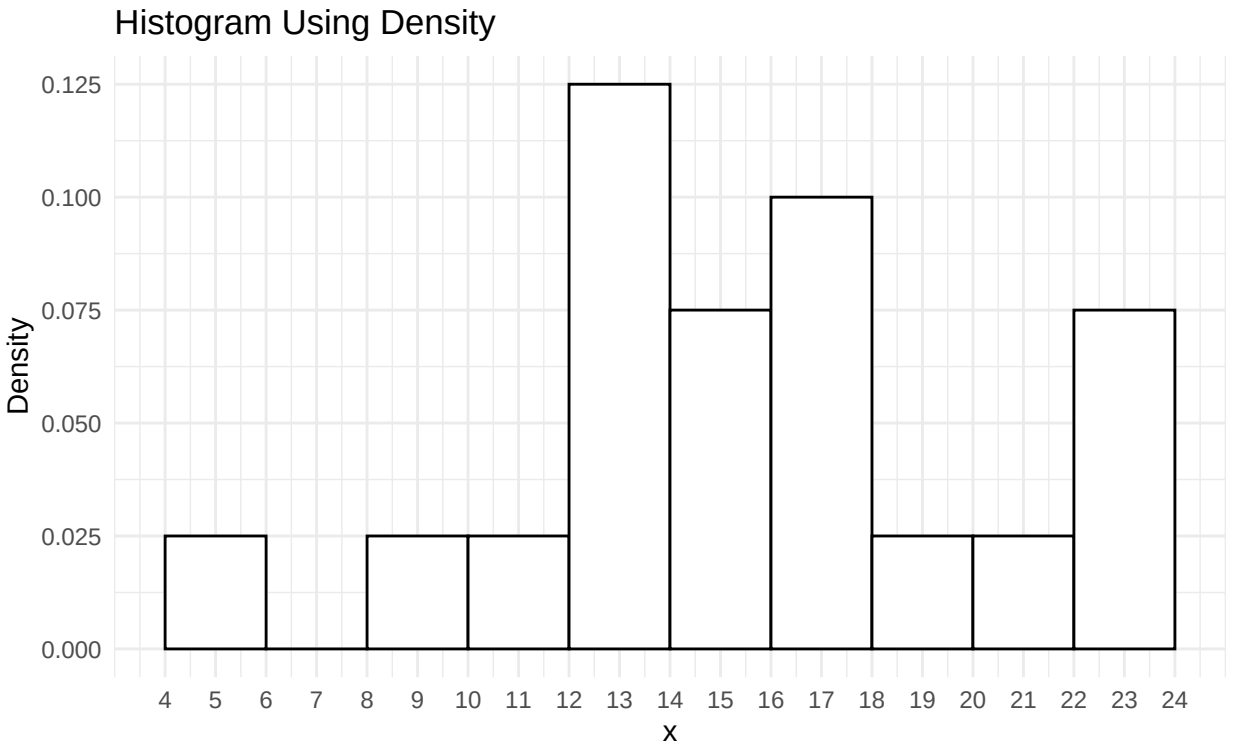
1. **Frequency** is simply the *number* of observations in a given interval. For example, there are 1 observations in the interval 10 to 12. Note that the sum of all of the frequencies will equal n.
2. **Relative frequency** is the *proportion* of observations in a given interval. It can be computed by dividing the frequency by n. For example, the relative frequency of the interval 10 to 12 is $1/20 = 0.05$. Note that the sum of the relative frequencies will equal 1.
3. **Density** is the *proportion of observations per unit of the number line*. It can be computed by dividing the relative frequency by the width of the interval. For example, the density of the interval 10 to 12 is $0.05/2 = 0.025$. Density is a bit more confusing than frequency or relative frequency, but it has some useful properties. One useful result of using density is that the *area* of a bar equals relative frequency —

i.e., the proportion of observations in that interval. Density is rarely used except in cases where it is desirable to make the intervals have unequal width.

Now a *histogram* is a graphical representation of the distribution of the quantitative variable that uses the intervals and either frequency, relative frequency, or density. A histogram is essentially a kind of bar graph. The position of each bar on the horizontal axis corresponds to the *interval* — i.e., the bar must cover the interval it represents. The height of each bar equals either frequency, relative frequency, or density. The figures below show the three kinds of histograms for the data shown earlier.







Note that the three histograms have the same *shape*. Only the scale of the vertical axis has changed. This is always true (provided that the intervals are all of equal width). So, the choice of frequency, relative frequency, or density is basically a matter of how we want to communicate how "often" observations fall into each interval.