1.2: Visual Representation of Data I - Categorical Variables

Suppose we have a population and variable in which we are interested. We get a sample, which could be large or small, and look at the values of the our variable for the individuals in that sample. We shall informally refer to this collection of values as a *dataset*.

In this section, we suppose also that the variable we are looking at is categorical. Then we can summarize the dataset by telling which categorical values did we see for the individuals in the sample, and how often we saw those values.

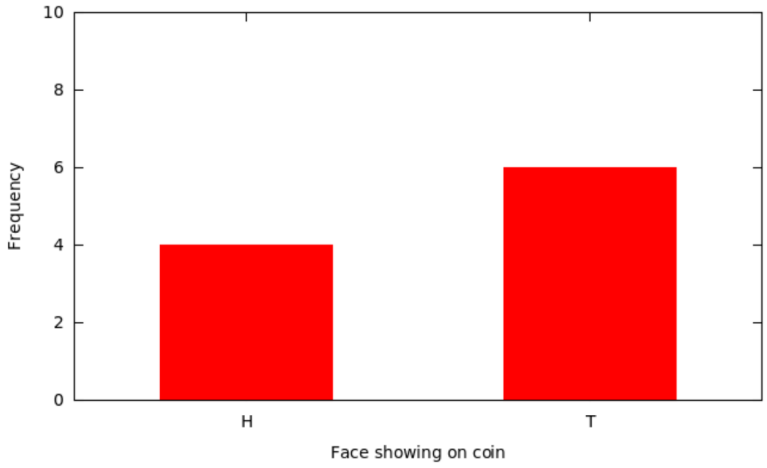
There are two ways we can make pictures of this information: *bar charts* and *pie charts*.

Bar Charts I: Frequency Charts

We can take the values which we saw for individuals in the sample along the x-axis of a graph, and over each such label make a box whose height indicates how many individuals had that value – the **frequency** of occurrence of that value.[def:frequency]

This is called a **bar chart**. As with all graphs, you should *always label all axes.* The x-axis will be labeled with some description of the variable in question, the y-axis label will always be "frequency" (or some synonym like "count" or "number of times").

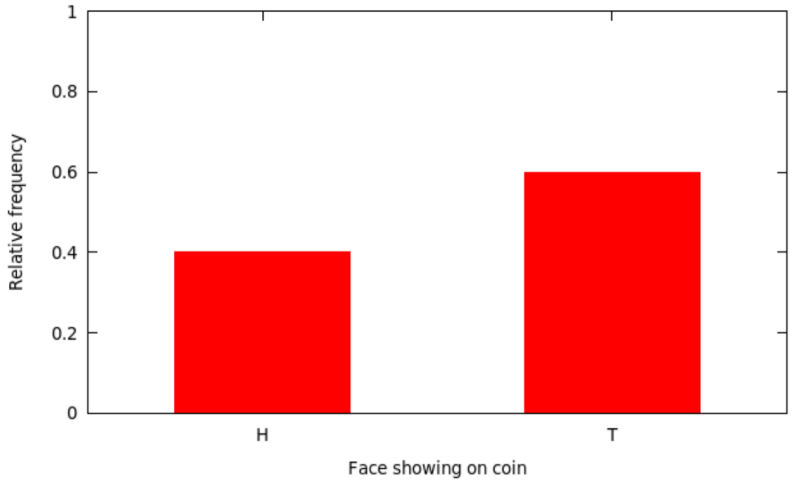
Example 1.2.1. In Example 1.1.7, suppose we took a sample of consisting of the next 10 flips of our coin. Suppose further that 4 of the flips came up heads – write it as "H" – and 6 came up tails, T. Then the corresponding bar chart would look like



Bar Charts II: Relative Frequency Charts

There is a variant of the above kind of bar chart which actually looks nearly the same but changes the labels on the y-axis. That is, instead of making the height of each bar be how many times each categorical value occurred, we could make it be *what fraction of the sample had that categorical value* – the **relative frequency**[def:relfreq]. This fraction is often displayed as a percentage.

Example 1.2.2. The relative frequency version of the above bar chart in Example 1.2.1 would look like



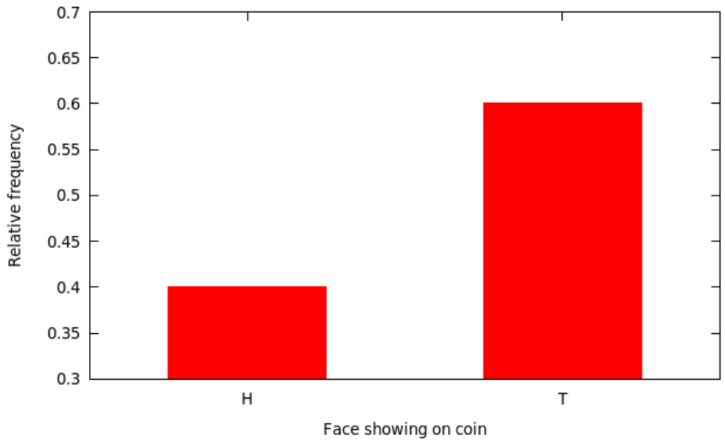
Bar Charts III: Cautions

Notice that with bar charts (of either frequency or relative frequency) the variable values along the x-axis *can appear in any order whatsoever*. This means that any conclusion you draw from looking at the bar chart must not depend upon that order. For example, it would be foolish to say that the graph in the above Example 1.2.1 "shows and increasing trend," since it would make just as much sense to put the bars in the other order and then "show a decreasing trend" – both are meaningless.

For relative frequency bar charts, in particular, note that the total of the heights of all the bars must be 1 (or 100%). If it is more, something is weird; if it is less, some data has been lost.

Finally, it makes sense for both kinds of bar charts for the y-axis to run from the logical minimum to maximum. For frequency charts, this means it should go from 0 to n (the sample size). For relative frequency charts, it should go from 0 to 1 (or 100%). Skimping on how much of this appropriate y-axis is used is a common trick to lie with statistics.

Example 1.2.3. The coin we looked at in Example 1.2.1 and Example 1.2.2 could well be a fair coin – it didn't show exactly half heads and half tails, but it was pretty close. Someone who was trying to claim, deceptively, that the coin was not fair might have shown only a portion of the y axis, as



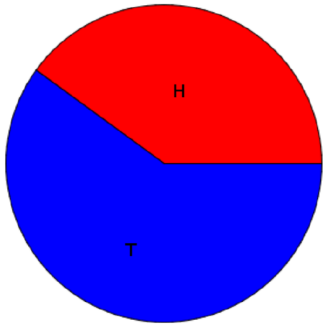
This is actually, in a strictly technical sense, a correct graph. But, looking at it, one might conclude that T seems to occur more than twice as often as H, so the coin is probably not fair ... until a careful examination of the y-axis shows that even though the bar for T is more than twice as high as the bar for H, that is an artifact of how much of the y-axis is being shown.

In summary, bar charts actually don't have all that much use in sophisticated statistics, but are extremely common in the popular press (and on web sites and so on).

Pie Charts

Another way to make a picture with categorical data is to use the fractions from a relative frequency bar chart, but not for the heights of bars, instead for the sizes of wedges of a pie.

Example 1.2.4. Here's a pie chart with the relative frequency data from Example 1.2.2.



Pie charts are widely used, but actually they are almost never a good choice. In fact, do an Internet search for the phrase "pie charts are bad" and there will be nearly 3000 hits. Many of the arguments are quite insightful.

When you see a pie chart, it is either an attempt (misguided, though) by someone to be folksy and friendly, or it is a sign that the author is quite unsophisticated with data visualization, or, worst of all, it might be a sign that the author is trying to use mathematical methods in a deceptive way.

In addition, all of the cautions we mentioned above for bar charts of categorical data apply, mostly in exactly the same way, for pie charts.
