A Useful Stemplot Example

Robert Kushler
Oakland University Department of Mathematics and Statistics (retired)

Abstract
A simple stemplot example, suitable for "live" construction during an introductory statistics lecture, is described. While presenting the example, a number of key issues that arise later in the course can be introduced and discussed.

Key Words: stemplots; teaching introductory statistics; statistical thinking

Stemplots (or "stem-and-leaf" plots) are a standard topic in introductory statistics courses, usually covered very early in the course, as one of several ways to make a graph of the distribution of a "continuous" variable. In simple cases, the last digit of the data value is the "leaf" and the remaining digits (usually including "leading zeros" to ensure that all values have the same total number of digits) are the "stem." The stems define the "bins" for what amounts to a histogram, and each "pile of leaves" is one of the bars. Variations involve using "split stems" to spread out the graph, or using the last two digits as the leaf in order to compress the graph (the goal in either case being to produce a "just right" picture of the distribution).

The key feature of this method is that it can be used to quickly produce a graph "by hand" from raw data (though ironically virtually all statistical software packages can be used to produce stemplots). For many years I have illustrated the manual process in my introductory statistics courses using the variable "birth day of month." Even in fairly large classes,the data can be quickly collected from the students "on the fly" while building the plot. More importantly, discussion of this example provides an opportunity to introduce several ideas that are covered more thoroughly later in the course.

The first step is defining the stems and leaves. There is really only one option in this case, but we do need a "leading zero" stem, and if the class size is large enough (say, over 40) the option of splitting each stem in two (leaves 0-4 and 5-9 in separate piles) is worth considering. If the class is large enough to require a "five split" (0-1, 2-3, 4-5, 6-7, and 8-9 in separate piles) then it is probably too large for the manual data collection process to be feasible.

Aside: when describing the stem-splitting options, I always ask the students why we can't create three or four separate piles. The answer of course is that using such splits would create a visually distorted graph with a non-uniform horizontal axis scale.

Before proceeding to collect the data, I pose two important questions. First, I ask the students what they expect to see when the graph is finished. They quickly come up with the idea that the "3" stem will have a much shorter pile than the others, which should be roughly equal. Upon further reflection (perhaps with a little prompting), we decide that the first pile should be shorter than the middle two piles since there is no "day zero." Finally, the fact that there is (usually) no February 29th means that the "2" stem should have a slightly shorter pile than the "1" stem – but at this point we are perhaps taking things too far. Tallying the days in a four year period yields 432/1461 = 0.296, 480/1461 = 0.329, 477/1461 = 0.326, and 72/1461 = 0.049 as the "expected" proportions on the 0, 1, 2, and 3 stems.

The second (and far more interesting and important) question is: *why* do we expect to see this pattern? The answer of course is that we think the day of the month on which you happen to be born is "random" – which in this case (but, I hasten to point out, not always) means "outcomes are equally likely." This is an example of statistical thinking, and we have just developed a statistical model for describing a pattern of real world variation.

Is the model plausible in this case? I would say yes. It is interesting to compare this case to the example "birth day of week." There is a well-documented "weekend effect" (perhaps due to obstetrician golf schedules), so the "equally likely" model is no doubt incorrect for day of week. (Aside: collecting "birth day of week" data in class is not feasible, though determining it is an interesting mathematical exercise, and a number of online resources for doing so are available.)

Another potentially useful comparison for discussion purposes is year of birth. A display of this variable for a random sample would mimic an age pyramid for the population, but the enrollees in a typical introductory statistics course are not a random sample (with respect to age – but for other purposes it may be possible to, in David Moore's felicitous phrase, "act as if" they are a random sample).

"Month of birth" could also be discussed in this context. Arguments against the "equally likely" (i.e., proportional to month length) model can be made (some more plausible than others). I recall seeing (a long time ago) an anecdote in Reader's Digest about the flurry of activity that occurs in maternity wards each March, in preparation for the influx of June brides.

After this discussion I proceed to collect the data, building the plot as I go. If the class is small enough, I take the time to sort each "pile of leaves," in order to demonstrate how quickly building a stemplot can produce a completely sorted data set – an otherwise extremely tedious process. I later remind them of this when I torture them (usually only once) with a homework exercise requiring manual caclulation of the median and quartiles.

Aside: it is often claimed that stemplots "preserve the data," but a key feature is lost. What is it? Answer: the order in which the values were obtained. This is often a crucial aspect of the data (but probably not in this case).

When the graph is finished, I ask the students whether they think it is consistent with our expectations (based on the model). After some discussion (during which students often reveal a tendency to overreact to random variation), I point out that this is an example of statistical inference, the logic of which will be developed later in the course, and that the specific technique for addressing this issue (consistency or lack thereof between a frequency distribution and a probability model) is covered in chapter XX of our textbook.

Finally, (if the stemplot has been sorted) I go through the stemplot and ask the students who were born on the same day of the month to call out the month of their birth. As you no doubt know, if the class size exceeds 22 I have a better than even chance of finding a birthday match, amazing the crowd. Either way, I promise to explain the logic of the birthday match probability calculation as one of the examples when we reach the dreaded chapter(s) on probability.

Coda: another topic I often discuss in connection with stemplots is rounding vs truncation. This issue arises when a "last digit leaf" plot will have far too many stems. Rather than using two digit leaves, the second to last digit can be used (admittedly this no longer preserves the data). Most students intuitively believe that rounding is "more accurate" than trucation, though of course the loss of accuracy is the same either way. Rounding is preferable when arithmetic operations are involved (e.g., the mean will be biased downwards if the data values are truncated). However, for the purpose of building a graph, the opposite holds - rounding will put some of the data values on the "wrong" stem. In addition, truncation (ignoring the last digit) is an easier mental operation than rounding.