Iterative Statistics

This library consists of simply classes used for tracking sample statistics from data sets that are not entirely observable all at once. An example use case is computing statistics from data stored across several files that cannot reasonably be loaded into memory concurrently.

Another practical use is computing statistics from a conditional subset of data. Instead of traversing data once to find observations that meet a condition, then making a second pass to compute statistics, the classes herein can track the necessary values on the first pass to report statistics once the traversal is complete.

Univariate Samples

For a univariate sample $\{x_i : i \in [1, n]\}$ the sample mean is computed as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

To compute this quantity all the is required is a count of the observations and a running sum of their values. Both are easily updated while traversing a data set in any fashion. We can similarly determine the variables needed for computing variance from its formula:

$$\begin{aligned}
\sigma^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \left({x_i}^2 - 2x_i\bar{x} + \bar{x}^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^{n} {x_i}^2 - 2\bar{x} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^{n} {x_i}^2 - 2 \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} x_i \right) + n\bar{x}^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^{n} {x_i}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right)
\end{aligned}$$

From this expansion we see that the variance requires three quantities in order to be computed: the count n, the sum of all observations x_i, and the sum of the squares of all observations x_i^2. Two of these values are already necessarily kept track of for computing the mean, so determining the variance from a sample requires storing only one additional floating point value.

