Estimating quantiles

Thomas Lumley
July 17, 2021

The $p$th quantile is defined as the value where the estimated cumulative distribution function is equal to $p$. As with quantiles in unweighted data, this definition only pins down the quantile to an interval between two observations, and a rule is needed to interpolate. As the help for the base R function `quantile` explains, even before considering sampling weights there are many possible rules.

Rules in the `svyquantile()` function can be divided into three classes

- Discrete rules, following types 1 to 3 in `quantile`
- Continuous rules, following types 4 to 9 in `quantile`
- A rule proposed by Shah & Vaish (2006) and used in some versions of SUDAAN

**Discrete rules**

These are based on the discrete empirical CDF that puts weight proportional to the weight $w_k$ on values $x_k$.

$$\hat{F}(x) = \frac{\sum \{x_i \leq x\} w_i}{\sum w_i}$$

**The mathematical inverse** The mathematical inverse $\hat{F}^{-1}(p)$ of the CDF is the smallest $x$ such that $F(x) \geq p$. This is rule `hf1` and `math` and in equally-weighted data gives the same answer as `type=1` in `quantile`.

**The primary-school median** The school definition of the median for an even number of observations is the average of the middle two observations. We extend this to say that the $p$th quantile is $q_{low} = \hat{F}^{-1}(p)$ if $\hat{F}(q_{low}) = p$ and otherwise is the the average of $\hat{F}^{-1}(p)$ and the next higher observation. This is `school` and `hf2` and is the same as `type=2` in `quantile`.

**Nearest even order statistic** The $p$th quantile is whichever of $\hat{F}^{-1}(p)$ and the next higher observation is at an even-numbered position when the distinct data values are sorted. This is `hf3` and is the same as `type=3` in `quantile`.

1
Continuous rules

These construct the empirical CDF as a piecewise-linear function and read off the quantile. They differ in the choice of points to interpolate. Hyndman & Fan describe these as interpolating the points \((p_k, x_k)\) where \(p_k\) is defined in terms of \(k\) and \(n\). For weighted use they have been redefined in terms of the cumulative weights \(C_k = \sum_{i \leq k} w_i\), the total weight \(C_n = \sum w_i\), and the weight \(w_k\) on the \(k\)th observation.

<table>
<thead>
<tr>
<th>(q) rule</th>
<th>Hyndman &amp; Fan</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>hf4</td>
<td>(p_k = k/n)</td>
<td>(p_k = C_k/C_n)</td>
</tr>
<tr>
<td>hf5</td>
<td>(p_k = (k - 0.5)/n)</td>
<td>(p_k = (C_k - w_k)/C_n)</td>
</tr>
<tr>
<td>hf6</td>
<td>(p_k = k/(n + 1))</td>
<td>(p_k = C_k/(C_n + w_n))</td>
</tr>
<tr>
<td>hf7</td>
<td>(p_k = (k - 1)/(n - 1))</td>
<td>(p_k = C_{k-1}/C_{n-1})</td>
</tr>
<tr>
<td>hf8</td>
<td>(p_k = (k - 1/3)/(n + 2/3))</td>
<td>(p_k = (C_k - w_k/3)/(C_n + w_n/3))</td>
</tr>
<tr>
<td>hf9</td>
<td>(p_k = (k - 3/8)/(n + 1/4))</td>
<td>(p_k = (C_k - 3w_k/8)/(C_n + w_n/4))</td>
</tr>
</tbody>
</table>

Shah & Vaish

This rule is related to hf6, but it is discrete and more complicated. First, define \(w^*_i = w_i/n/C_n\), so that \(w^*_i\) sum to the sample size rather than the population size, and \(C^*_k\) as partial sums of \(w^*_k\). Now define the estimated CDF by

\[
\hat{F}(x_k) = \frac{1}{n+1} \left( \frac{C^*_k + 1/2 - w_k/2}{w_k/2} \right)
\]

and take \(\hat{F}^{-1}(p)\) as the \(p\)th quantile.

Other options

It would be possible to redefine all the continuous estimators in terms of \(w^*\), so that type 8, for example, would use

\[
p_k = \frac{(C^*_k - 1/3)/(C^*_n + 2/3)}{w_k/3}/(C^*_n + 2/3).
\]

Or a compromise, eg using \(w^*_k\) in the numerator and numbers in the denominator, such as

\[
p_k = \frac{(C^*_k - w^*_k/3)/(C^*_n + w^*_n/3)}{w_k}/(C^*_n + w^*_n/4).
\]

Comparing these would be a worthwhile... an interesting... a research question for simulation.