Methods Used for Estimating Percentiles in EdSurvey

Developed by Paul Bailey and Michael Cohen[*][†]

April 30, 2018

The EdSurvey package gives users functions to efficiently analyze education survey data. This vignette describes the methods used to estimate percentiles in EdSurvey.

Note:

New users looking for an introduction and overview of EdSurvey should see *Using EdSurvey to Analyze NCES Data: An Illustration of Analyzing NAEP Primer.*[1]

Estimating Percentiles

Unweighted percentiles (P) are estimated according to the method recommended by Hyndman and Fan (1996), the median unbiased estimator of order $o(n^{-1/2})$. Hyndman and Fan wrote, “Reiss also states that the resulting sample quantile is optimal in the class of all estimators that are median unbiased of order $o(n^{-1/2})$ and equivariant under translations” (1996, pages 363–364, citing Reiss [1989]).[2]

This method uses the ordered pairs of a (percentile, x-value) (p_k, x_k) where x_k is the value of the kth smallest value of x and $p_k = \frac{k-1/3}{n+1/3}$. A linear interpolation between the two closed points to a percentile of the required percentile (p) is then calculated. (Because these ordered pairs do not span [0,1] when n is finite, a value of $(0, x_1)$ is added to the beginning of the list and $(1, x_n)$ to the end.)

To perform the linear interpolation, the value k is identified, such that $p_k \leq p$ and $p_{k+1} > p$. Then, γ is defined as

$$\gamma = (p - p_k)/(p_{k+1} - p_k) \tag{1}$$

and the percentile is defined as

$$P = (1 - \gamma)x_k + \gamma x_{k+1} \tag{2}$$

When weighted percentiles are estimated, the weighted equivalent is used, so that the value of k is again found to meet the same criterion ($p_k \leq p$ and $p_{k+1} > p$), but now p_k is defined as follows:

$$p_k = \frac{\frac{n}{W}\left(\sum_{i=1}^{k} w_i\right) - 1/3}{n + 1/3} \tag{3}$$

[*] This publication was prepared for NCES under Contract No. ED-IES-12-D-0002 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

[†] The authors would like to thank Dan Sherman for reviewing this document.

[1] This document is available online at http://www.air.org/sites/default/files/EdSurvey.pdf.

[2] In R’s `quantile` function, this is the method selected by setting the `type` argument to eight. Hyndman and Fan (1996) recommended this method because it has a motivation in estimation (it is median unbiased, regardless of the distribution) while meeting five of the six desirable properties for percentile estimators.

where w_i is the weight associated with the i^{th} smallest value of x and $W = \sum_{i=1}^{n} w_i$.

The same formula for the percentile is used.

When the values of x are a set of plausible values, the mean of the percentile over the many plausible values is used.

$$P = \frac{1}{m}\sum_{j=1}^{m} p_j \tag{4}$$

where m is the number of plausible values, and p_j is the estimated percentile for the jth set of plausible values.

Estimation of Dispersion Parameters for Percentiles

The variance estimation of percentiles is more complicated than the variance estimation of means because they are defined on only a finite space (from the minimum to the maximum) and are not differentiable.[3] Consequently, `EdSurvey` generates both standard error estimates as well as a confidence interval.

Estimating the Standard Errors for Percentiles

The standard error estimates, as recommended by Johnson and Rust (1992), use the jackknife variance estimator in the statistics vignette sections “Estimation of Standard Errors of Weighted Means When Plausible Values Are Not Present, Using the Jackknife Method” or “Estimation of Standard Errors of Weighted Means When Plausible Values Are Present, Using the Jackknife Method.”

Estimating the Confidence Intervals for Percentiles

A confidence interval can be built using the jackknife standard errors. This method should work well in most cases but does fully acknowledge that percentiles are not defined to span the entire space of $\mathbb{R}$ and that their best confidence interval may not be symmetric.

A second method, which does allow for asymmetric confidence intervals that fall entirely in the range $[x_1, x_n]$, follows Woodruff (1952). When the Pth percentile is estimated as $\hat{p}$, the variance (v) of the proportion of the population above or below $\hat{p}$ is estimated, and this variance is used to construct a percentile, in the proportion space, according to

$$[P + v^{\frac{1}{2}} \cdot t_{\frac{\alpha}{2}}, \, P + v^{\frac{1}{2}} \cdot t_{1-\frac{\alpha}{2}}] \tag{5}$$

These proportions are then percentiles, and their level can be transformed back to levels of the variable in the same way that the percentile was estimated.

When the variable in question does not contain plausible values, the variance of the proportion of the population below the estimated percentile is estimated according to the statistics vignette sections “Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Not Present, Using the Jackknife Method” or “Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Not Present,

[3] Strictly speaking, percentiles are piecewise differentiable, when estimated in the way described in this document. However, these formulas were intended to generate percentiles, not reliable estimates of the first derivative of the percentiles.

Using the Taylor Series Method." In this case, $\mathcal{A}$ is the set of all values less than the estimated value for the percentile, and $\mathcal{U}$ is the set of all values.

When plausible values are present, the set $\mathcal{A}$ is based on the estimated value for the percentile *after averaging across the plausible values*. Here, membership in $\mathcal{A}$ is dependent on plausible values, so the variance is estimated according to the sections "Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Present, Using the Jackknife Method" or "Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Present, Using the Taylor Series Method."

References

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, *50*, 361–365.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Education Statistics*, *17*(2), 175–190.

Lee, M. D., Bailey, P. D., Emad, A., Zhang, T., Nguyen, T. M., & Yu, J. (2018). *Using EdSurvey to analyze NCES data: An illustration of analyzing NAEP primer.* Washington, DC: American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/EdSurvey.pdf

Reiss, R.-D. (1989). *Approximate distributions of order statistics: with applications to nonparametric statistics*. Springer-Verlag.

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, *47*(260), 635–646.