perccalc: An R package for estimating percentiles from categorical variables

Jorge Cimentada[1]

1 Laboratory of Digital and Computational Demography, Max Planck Institute of Demographic Research (MPIDR)

Summary

Social science research makes extensive use of categorical variables. This means that most variables in model definitions are a combination of categorical and ordered categorical variables, which sometimes are proxies of continuous variables such as income or years of education. The seriousness of this phenomena can be best exemplified by the surge and usage of techniques tailored specifically for this type of analysis in social science research (Agresti, 2007, 2010).

In particular, educational research, where there's a maturing literature on calculating inequality gaps, categorical data are essential for estimating inequality. For example, the income of a person is often asked in income brackets rather than the exact amount of money; researchers would prefer the exact amount but to avoid non-response accumulation and privacy concerns, income brackets are a partial solution. This solution gives the income information of respondents but at the same time in a limited fashion given that we cannot estimate traditional statistics such as the differences of percentiles from the income brackets. One example of this is calculating the gap in cognitive abilities between the top (e.g 90th percentiles) and bottom (e.g 10th percentiles) groups in the income distribution.

`perccalc` is a direct implementation of the theoretical work of Reardon (2011) where it is possible to estimate the difference between two percentiles from an ordered categorical variable. More concretely, by specifying an ordered categorical variable and a continuous variable, this method can estimate differences in the continuous variable between percentiles of the ordered categorical variable. This bring forth a relevant strategy to contrast ordered categorical variables which usually have alternative continuous measures to the percentiles of the continuous measures. Moreover, this opens an avenue for calculating percentile distributions and percentile differences for ordered categorical variables which don't necessarily have an alternative continuous measure such as job occupation classifications; one relevant example being the classification from Erikson, Goldthorpe, & Portocarero (1979).

Recently, this method has been growing in usage in education research (Bassok, Finch, Lee, Reardon, & Waldfogel, 2016; Chmielewski & Reardon, 2016; Reardon, 2011; Reardon & Portilla, 2016), yet this technique is not limited to this field alone and can be used essentially in any context where percentiles of ordered categorical variables are of interest. One example where this would provide useful would be in medicine based research, where demographic characteristics such as education categories are common factors for looking at differences between groups.

The field of computational categorical data analysis has a long history in R with packages addressing small-area estimation for categorical variables (Boonstra, 2012), missing data imputation (van Buuren & Groothuis-Oudshoorn, 2011) and standard generalized models for ordinal data (Christensen, 2019). The `qualvar` package (Gombin, 2018) is one attempt to focus not on the modelling of categorical variable but rather on the properties of such variables

to calculate variation in categorical variables. Yet despite the popularity of categorical-based methods, there is still not an official software package that reliably implements and tests Reardon's method in the **R** programming language (R Core Team, 2019); nor in any other programming language, that I'm aware of.

Currently, `perccalc` implements:

- Calculating differences in a continuous variable relative to the percentiles of an ordered categorical variable
- Calculating values for a continuous variable relative to the percentiles of an ordered categorical variable (values of a continuous variable for the 1th, 10th, 20th, …, 100th percentile of the ordered categorical variable)
- Weight-adjusted estimations for all percentile calculations
- Provides uncertainty estimates for all calculations which allows the user to produce uncertainty intervals or propagate further calculations with these uncertainty coefficients

`perccalc` offers flexibility and reliability for estimating any number of percentile differences for ordered categorical variables as well as the distribution of percentiles values for an ordered categorical variable. Moreover, it provides the standard errors for the estimation which can be used to construct uncertainty intervals. This full-featured implementation offers a reliable software to use in serious peer-review research. Researchers can trust this implementation as an accurate representation given that it has been built by testing it to decimal accuracy to the theoretical model of Reardon (2011); these tests are continually checked on a weekly basis making the package particularly reliable.

The major features (including examples addressing real world problems) of `perccalc` are shown in a series of vignettes in the package's website (https://cimentadaj.github.io/perccalc/), where there is a direct implementation that matches Reardon (2011)'s initial implementation. Additionally, the package is hosted on it's own open source repository on Github (https://github.com/cimentadaj/perccalc/) and on the official CRAN repository (https://cran.r-project.org/web/packages/perccalc/index.html)

Acknowledgements

References

Agresti, A. (2007). An introduction to categorical data analysis. John Wiley & Sons, Inc. doi:10.1002/0470114754

Agresti, A. (2010). *Analysis of ordinal categorical data* (Vol. 656). John Wiley & Sons.

Bassok, D., Finch, J. E., Lee, R., Reardon, S. F., & Waldfogel, J. (2016). Socioeconomic gaps in early childhood experiences: 1998 to 2010. *AERA Open*, *2*(3), 2332858416653924. doi:https://doi.org/10.1177/2332858416653924

Boonstra, H. J. (2012). *Hbsae: Hierarchical Bayesian small area estimation*. Retrieved from https://CRAN.R-project.org/package=hbsae

Chmielewski, A. K., & Reardon, S. F. (2016). Patterns of cross-national variation in the association between income and academic achievement. *AERA Open*, *2*(3), 2332858416649593. doi:https://doi.org/10.1177/2332858416649593

Christensen, R. H. B. (2019). Ordinal—regression models for ordinal data. Retrieved from http://www.cran.r-project.org/package=ordinal/

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology*, *30*(4), 415–441. doi:https://doi.org/10.2307/589632

Gombin, J. (2018). *Qualvar: Implements indices of qualitative variation proposed by Wilcox (1973)*. Retrieved from https://CRAN.R-project.org/package=qualvar

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither opportunity*, 91–116.

Reardon, S. F., & Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open*, *2*(3), 2332858416657343. doi:https://doi.org/10.1177/2332858416657343

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. doi:https://doi.org/10.18637/jss.v045.i03