The Journal of Open Source Software

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)

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

Social science research is hampered by the 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, 2010; Agresti & Kateri, 2011).

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 amout 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

Below I introduce the reader to one simplified example of `perccalc`'s capabilities in a real world scenario.

Example

For this example, we will calculate the percentile difference in Mathematics test scores based on the education of the parents for several countries using the PISA 2006 and PISA 2012 datasets. Let's load `perccalc` with the packages `dplyr` (Wickham et al., 2019), `tidyr` (Wickham & Henry, 2019) and `ggplot2` (Wickham, 2016) for wrangling and visualizing the data:

```
library(perccalc)
library(tidyr)
library(ggplot2)
library(dplyr)
```

`percalc` automatically loads `pisa_2012` and `pisa_2006` which are two datasets with all the information that we need. These two datasets have data for Estonia, Germany and Spain and contain the average test score in Mathematics for each student and their father's education as measured by the international ISCED classification. The first thing we have to do is make sure the categories are ordered.

```
order_edu <- c("None", "ISCED 1", "ISCED 2",
               "ISCED 3A, ISCED 4", "ISCED 3B, C",
               "ISCED 5A, 6", "ISCED 5B")

# Make ordered categories of our categorical variables
pisa_2006 <-
  pisa_2006 %>%
  mutate(father_edu = factor(father_edu, levels = order_edu, ordered = TRUE))

pisa_2012 <-
```

```
  pisa_2012 %>%
  mutate(father_edu = factor(father_edu, levels = order_edu, ordered = TRUE))

# Merge them together
pisa <- rbind(pisa_2006, pisa_2012)
```

Once the categories are ordered, `perc_diff` can calculate the percentile difference between the 90th and 10th percentile for the complete sample. For example:

```
perc_diff(data_model = pisa,
          categorical_var = father_edu,
          continuous_var = avg_math,
          percentiles = c(90, 10)
          )
```

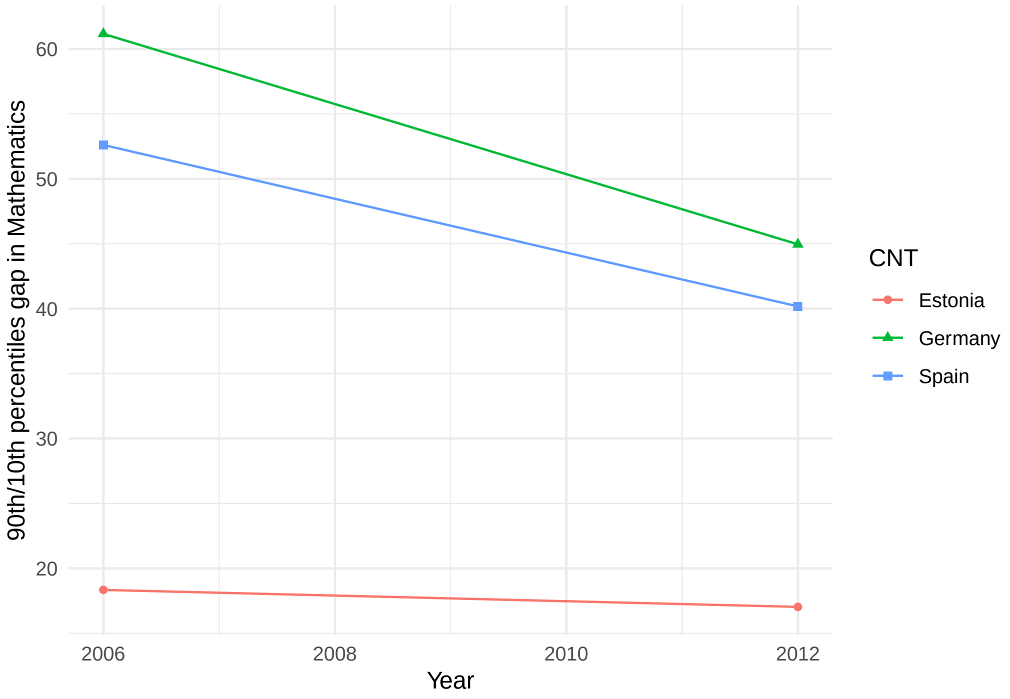
```
## difference         se
##   45.07312   14.18308
```

This means that the difference in Mathematics test scores between the 90th and 10th percentile of father's education is 45 points with a standard error of 14 points. We can extrapolate this example for each country separately using `perc_diff_df` which returns a data frame:

```
cnt_diff <-
  pisa %>%
  nest(data = c(-year, -CNT)) %>%
  mutate(
    # Look through each year/cnt combination and apply perc_diff_df
    edu_diff = lapply(data, function(x) {
      perc_diff_df(data_model = x,
                   categorical_var = father_edu,
                   continuous_var = avg_math,
                   percentiles = c(90, 10))
    })
  ) %>%
  select(-data) %>%
  unnest(edu_diff)

title_y_axis <- "90th/10th percentiles gap in Mathematics"

# Plot results
cnt_diff %>%
  ggplot(aes(year, difference, color = CNT, shape = CNT)) +
  geom_point() +
  geom_line() +
  theme_minimal() +
  scale_y_continuous(name = title_y_axis) +
  scale_x_continuous(name = "Year")
```



It looks like Estonia has a much smaller achievement gap relative to Spain and Germany but also note that both Germany and Spain have been decreasing their inequality.

In contrast, `perc_dist` calculates the distribution of the percentiles which is useful for comparing more finegrained percentiles rather than differences.

```
perc_dist(data_model = pisa,
          categorical_var = father_edu,
          continuous_var =  avg_math
          )
```

```
## # A tibble: 100 x 3
##    percentile estimate std.error
##         <int>    <dbl>     <dbl>
##  1          1    0.853      1.64
##  2          2    1.74       3.20
##  3          3    2.65       4.69
##  4          4    3.60       6.10
##  5          5    4.57       7.45
##  6          6    5.57       8.72
##  7          7    6.59       9.93
##  8          8    7.64      11.1
##  9          9    8.71      12.2
## 10         10    9.80      13.2
## # ... with 90 more rows
```

Here we get the complete father's education percentile distribution with the test score values in Mathematics for each percentile. This can be easily scaled to all country/year combinations with our previous code:

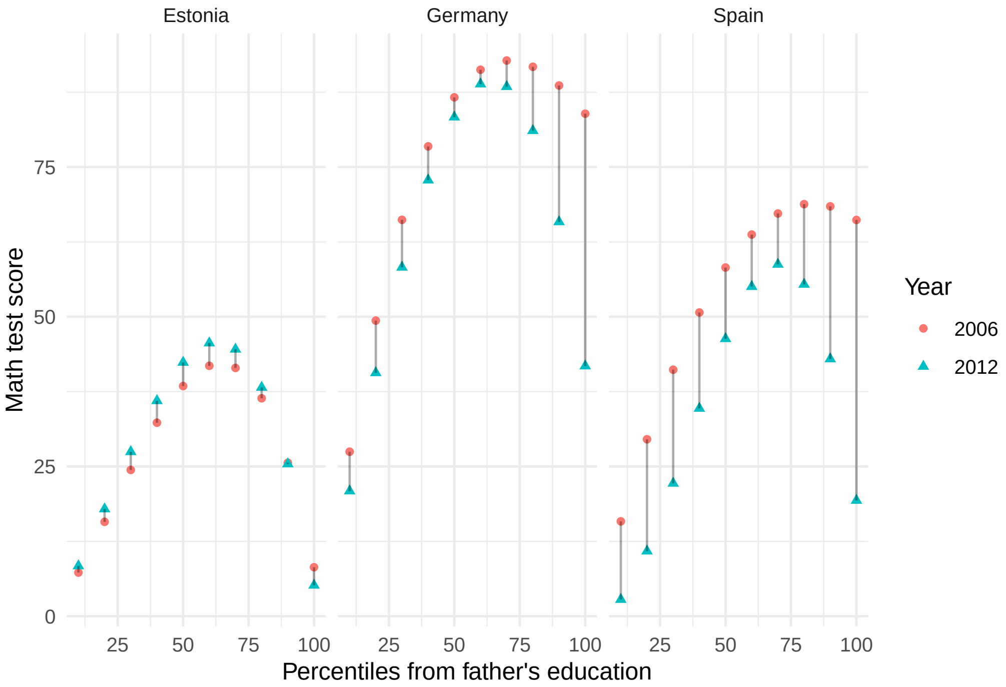
```
cnt_dist <-
  pisa %>%
  nest(data = c(-year, -CNT)) %>%
  mutate(
    edu_diff = lapply(data, function(x) perc_dist(x, father_edu, avg_math))
  ) %>%
  select(-data) %>%
  unnest(edu_diff)

cnt_dist
```

```
## # A tibble: 600 x 5
##     year CNT     percentile estimate std.error
##    <dbl> <chr>        <int>    <dbl>     <dbl>
##  1  2006 Estonia          1    0.647     0.388
##  2  2006 Estonia          2    1.31      0.762
##  3  2006 Estonia          3    2.00      1.12
##  4  2006 Estonia          4    2.71      1.47
##  5  2006 Estonia          5    3.43      1.81
##  6  2006 Estonia          6    4.17      2.13
##  7  2006 Estonia          7    4.93      2.44
##  8  2006 Estonia          8    5.70      2.74
##  9  2006 Estonia          9    6.48      3.02
## 10  2006 Estonia         10    7.28      3.30
## # ... with 590 more rows
```

Let's limit the distribution only to the 10th, 20th, 30th… 100th percentile and compare for country/years:

```
cnt_dist %>%
  mutate(year = as.character(year)) %>%
  filter(percentile %in% seq(0, 100, by = 10)) %>%
  ggplot(aes(x = percentile,
             y = estimate,
             color = year,
             shape = year,
             group = percentile)) +
  geom_point() +
  geom_line(color = "black", alpha = 1/3) +
  scale_y_continuous(name = "Math test score") +
  scale_x_continuous(name = "Percentiles from father's education") +
  scale_color_discrete(name = "Year") +
  scale_shape_discrete(name = "Year") +
  facet_wrap(~ CNT) +
  theme_minimal()
```



Here the dots indicate the year 2006 and the triangles year 2012, the grey line between them indicates the change over time. Here we can see that although Germany and Spain are decreasing (as we saw in the plot before), the composition of the change is very different: Spain's decrease is big all around the distribution whereas Germany's concentrates on the top percentiles.

In summary, `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 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 which matches Reardon (2011)'s initial implementation. Additionally, the package is hoted 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

I acknowledge contributions from Sean Reardon in the writing of this software both in the form of recommendations and of the theoretical underpinnings of the implementation.

References

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

Agresti, A., & Kateri, M. (2011). *Categorical data analysis*. Springer.

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.

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. Retrieved from https://www.jstatsoft.org/v45/i03/

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr