LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Social Policies and Distributional Outcomes
in a Changing Britain

distoutc and svydistoutc: Help file to accompany Stata programmes for undertaking distributional analysis of categorical outcome variables

Programme developed for the Social Policies and Distributional Outcomes Programme

Eleni Karagiannaki

SPDO research note 4

January 2022

CASE
Centre for Analysis of Social Exclusion
Research at LSE

Nuffield Foundation

The Social Policies and Distributional Outcomes in a Changing Britain (SPDO) research programme

The central objective of the Social Policies and Distributional Outcomes in a Changing Britain (SPDO) research programme is to provide an authoritative, independent, rigorous and in-depth evidence base on social policies and distributional outcomes in 21st century Britain. The research programme addresses the central question "*what progress has been made in addressing inequalities through social policy making?"* It is ambitious and comprehensive in scope, combining in-depth quantitative analysis of trends in distributional outcomes across multidimensional domains (living standards, employment, health and care, and physical security and security) by different characteristics (age, gender, disability, ethnicity/nationality/migration status, socio-economic group and area), with detailed social policy analysis ten major social policy areas (social security/general housing, health, social care, education, higher education, early years, employment, physical safety and security, homelessness/complex needs, social mobility). The research programme updates our previous Social Policy in a Cold Climate research programme and combines analysis of the current period (2015-2020) with broader reflection on the changing nature of social policies and distributional outcomes over the 21st century. Further details and research papers from the programme are available on the SPDO website, (http://sticerd.lse.ac.uk/case/_new/research/spdo/default.asp).

Abstract

The programmes estimate and automatically save in the current directory two excel files distributional statistics for any categorical variable *varname* (e.g. employment status, poverty status etc) overall and for each of the groups identified by each variable included in the variable list *varlist* (typical variables that can be included in the variable list include gender, age group, socio-economic status). Specifying the bygroup (*groupvar*) option performs the distributional analysis separately for each subgroup in *groupvar*. Both cross-sectional and over time change analysis is undertaken by the programme. The cross-sectional analysis produces estimates of the proportion of the sample that falls in each category identified by the categorical variable *varname* for each of the group identified by the variable in the *varlist* and for each year included in the dataset, as well as differences in the relevant statistics between the different groups relative to the reference group for each variable in the *varlist* (along with their respective standard errors and p-values). The over-time analysis produces estimates of the change in the proportion of the population that falls in each category defined by the categorical variable for each year t relative to the previous year t-1 (along with the standard error and p-value of the changes). The over-time analysis also produces estimates (along with their standard errors and the corresponding p-values) of the across groups differences in the change in the proportion of the population that falls in each category defined by the categorical variable *varname* relative to the reference category for each of the subpopulations identified by the categories in each variable in the *varlist*. The programme also allows analysis to be undertaken by subpopulation groups allowing intersectional research. Both a non-survey design and survey design versions of the programme are available (distoutc and svdistouc respectively).

Acknowledgements

The project has been funded by the Nuffield Foundation and the authors would like to thank the Foundation as well as the many people who provided comments on an earlier draft of this paper.

The Nuffield Foundation is an endowed charitable trust that aims to improve social well-being in the widest sense. It funds research and innovation in education and social policy and also works to build capacity in education, science and social science research. The Nuffield Foundation has funded this project, but the views expressed are those of the authors and not necessarily those of the Foundation. More information is available at www.nuffieldfoundation.org.

Responsibilities for errors remains with the author.

Note: This research note is a help guide to accompany the Stata distoutc and svydistoutc .do files. If you use this programme, please acknowledge this and reference this programme as "name of programme" authored by Eleni Karagiannaki. If you identify any errors or believe some aspect requires further clarification in this guide, please contact Eleni Karagiannaki (e.karagiannaki@lse.ac.uk).

Help for distoutc, svydistoutc Eleni Karagiannaki (December 2019)

Syntax

distoutc *varname varlist [, bygroup(groupvar)]* **pweights and aweights** are allowed**; see help weights.**

svydistoutc *varname varlist [, bygroup(groupvar)]* **pweights and aweights** are allowed; **see help weights.**

Important notes

svydistoutc:
strata and psu variables have to be named strata and psu respectively.

distoutc and svydistoutc:

The dataset must contain a variable named year to indicate year.
If the analysis does not involve an over-time analysis (or the dataset does not include year) the variable must be created by the user.

Options

bygroup (groupvar) requests inequality decompositions by population subgroup, with subgroup membership summarized by groupvar.

Description

distoutc estimates and automatically saves in excel file (under the name *outcome_varname*.xls) various distributional statistics of the categorical variable *varname* overall and for each of the groups identified by each variable in the *varlist* (typical variables which can be included in the *varlist* include gender, age group, socio-economic status). Specifying the bygroup (*groupvar*) option performs the distributional analysis separately for each subgroup in *groupvar*.

svydistoutc produces the same results but correcting the standard errors of the estimates for survey design effects (the user must strata and psu variables have to be named strata and psu) in the dataset. Both cross-sectional and over time analysis is undertaken by both programmes. The estimates produced by the cross-sectional and over time analysis are described below.

Cross-sectional distributional analysis estimates: **(svy)*distoutc*** produces estimates of the proportion of the sample that falls in each category identified by the categorical variable *varname* (with the different categories specified by *varname`j' where j=1..J* are the J categories of the categorical variable *varname)* for each of the group identified by the variable in the *varlist* and for each year included in the dataset - or for the years specified by the [**if** *exp*]. The standard errors of the proportion estimates and their corresponding p-values are also included in the saved excel file which holds the results.

Differences in the proportion of the sample that falls in each category of the *varname* for each group defined by each variable in the *varlist* relative to the reference category are also produced along with their standard errors and the corresponding p-values. The proportions of different subgroups falling in each category are estimated using the over(*varlist* [, nolabel]) option, which specifies the estimates to be computed for each of the subpopulations as identified by the different values of each variable in *varlist*. We use the lincom and test stata commands to for the statistical testing of the across group differences.

Over-time distributional analysis estimates: In addition to the cross-sectional estimates described above **(svy)*distoutc*** produces estimates of the change in the proportion of the population that falls in each category defined by the categorical variable *varname* for each year *t* relative to the previous year *t-1*. Estimates are produced both for the overall population and for each of the groups identified by the categories of each variable in the *varlist*. The standard errors and the p-values of the estimates of the change are also included in the results. The user can specify the years to be included in the analysis using the [**if** exp] option in order to examine patterns of change for years not covered by the default. The estimates of change in the proportions in each category of the *varname* for each subpopulation is undertaken using the **over**(year[, nolabel]) option. In distroutc the over-time change estimates for each group is derived restricting the estimation sample using the [**if** exp] option whereas in svydistroutc subpopulation analysis is undertaken using the **subpop**() option.

Across groups differences in the change in the proportion of the population that falls in each category defined by the categorical variable *varname* relative to the reference category for each of the subpopulations identified by the categories in each variable in the *varlist* are also produced along with their standard errors and the corresponding p-values. The difference in the change in the proportion in each category is estimated using a linear probability model with the probability of falling in each category j of the outcome variable varname as a dependent variable and a dummy for year (taking the value of 1 for year t and 0 for year t-1), a set of dummies indicating the different categories of each variable in the varlist and their interaction with year.

Estimation commands used in the programme

Cross-sectional proportion estimates and their standard error and p-values for the subpopulations are identified by the different values of the variables `var’

distoutc

```
        proportion varname`l' if year==`y' [pw=`wgt'],over (`var')
```

svydistoutc

```
        svy: proportion varname`j'  if year==`y',over (`var')
```

Note: varname`j' is a set of dummy variables indicating the different categories of the *varname*

Cross-sectional differences in proportions across the subpopulations identified by the different values of each variable in the *varlist* (relative to the reference category of each variable)

distoutc

```
proportion varname`j' if year==`y' [pw=`wgt'],over(`var')
lincom [_prop_2]`i'-[_prop_2]1
```

svydistoutc

```
svy: proportion varname `j'  if year==`y',over(`var',nolab) nolab
lincom [_prop_2]`i'-[_prop_2]1
```

Over time change estimates

distoutc

```
proportion varname`j' if `var'==`i' [pw=`wgt'],over(year)
lincom [_prop_2]`y'-[_prop_2]1
```

svydistoutc

```
     svy, subpop(subpop): proportion  varname`j',over(year)
     lincom [_prop_2]`y'-[_prop_2]1
```

Over time across group difference in the change

distoutc

```
reg varname`j' i.year   i.`var’  i.year#i.`var’  if  (year==`y'-1|  year==`y')
[pw=`wgt']
```

svydistoutc

```
        svy:reg varname`l' i.year i.`var’ i.year#i.`var’ if
        (year==`y'-1| year==`y')
```

Results saved in excel spreadsheets

Cross-sectional e*stimates*: *varname*`j’pY[year] indicates the proportion of each group defined by each variables in the varlist that falls in each category j (where j=1…J) of the outcome variable varname in each year [year]

***Standard errors of the cross-sectional estimates**:* *varname*`j’seY[year] indicates the standard error of the proportion of each group defined by each variables in the varlist that falls in each category j (where j=1…J) of the outcome variable varname in each year [year]

***P-values of the cross-sectional estimates**:* *varname*`j’pvY[year] indicates the p-values of the proportion of each group defined by each variables in the varlist that falls in each category j (where j=1…J) of the outcome variable varname in each year [year]

Across group difference estimate: d*varname*`j’pY[year] indicates the difference in the proportion of each group defined by each variable in the *varlist* relative to the reference category that falls in each category j (where j=1…J) of the outcome variable *varname* in each year [year]

Standard errors of the across group difference estimate: dvarname`j’seY[year] where j=1…J indicates the standard error of the difference in the proportion of each group defined by each variable in the *varlist* relative to the reference category that falls in each category j (where j=1…J) of the outcome variable *varname* in each year [year]

***P-values of the across group difference estimate**:* dvarname`j’pvY[year] indicates the p-value of the difference in the proportion of each group defined by each variable in the *varlist* relative to the reference category that falls in each category j (where j=1…J) of the outcome variable *varname* in each year [year]

Change in the estimate from previous year: c*varname*`j’pY[year] indicates the change in the proportion of each group defined by each variables in the *varlist* that falls in each category j (where j=1…J) of the outcome variable *varname*

Standard errors of the change from previous year in the estimates: c*varname*`j’seY[year] indicates the standard errors of the change in the proportion of each group defined by each variables in the *varlist* that falls in each category j (where j=1…J) of the outcome variable *varname*

P-values of the change from previous year in the estimates: cvarname`j’pvY[year] indicates the p-values of the change in the proportion of each group defined by each variables in the *varlist* that falls in each category j (where j=1…J) of the outcome variable *varname*

Across group difference in the change from previous year estimate: cd*varname*`j’pY[year] indicates the difference in the change in the proportion of each group defined by each variable in the *varlist* that falls in each category j (where j=1…J) of the outcome variable *varname* relative to the reference category of each variable in the *varlist*

Standard errors of the across group difference in the change from the previous estimate: cdvarname`j’pY[year] indicates the standard error of the difference in the change in the proportion of each group defined by each variable in the *varlist* that falls in each category j (where j=1…J) of the outcome variable *varname* relative to the reference category of each variable in the *varlist*

P-values of the across group difference estimate: cd*varname*`j’pY[year] indicates the p-value of the difference in the change in the proportion of each group defined by each variables in the *varlist* that falls in each category j (where j=1…J) of the outcome variable *varname* relative to the reference category of each variable in the *varlist.*

Explanation of the excel files

After you run the programme, you will have two datasets saved in the current directory named "*varname*_Layout1.xls **"** and **"***varname*_Layout2.xls**"**.

Layout 1: Results are organised by year

*Name of excel file: varname*_Layout1.xls

Description of layout1: Cross-sectional results for each year are saved in different excel worksheets under the name Cross-sectional analysis Y[year]. Over time change estimates for each year relative to the previous year are saved under the name Y[year] Change from base year.

Layout 2: Results are organised by categories of the *varname*

Name of the excel file: varname_Layout2.xls:

Description of layout2: Cross-sectional and over time estimates for each category of the *varname* are saved in different excel worksheets under the name *varname*`j' where j=1…J indicate the categories of the categorical variable *varname*.