THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Social Policies and Distributional Outcomes

in a Changing Britain

distout and svydistout: Help file to accompany Stata programmes for undertaking distributional analysis of continuous outcome variables

Programme developed for the Social Policies and Distributional Outcomes Programme

Eleni Karagiannaki

SPDO research note 1

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 distcout and svydistout programmes estimate and automatically save in the current directory two excel files which store a range of distributional statistics for any continuous variable (e.g. earnings, equivalised household disposable income) overall and for each of the groups specified by the variable include in the variable list (typical variables that 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. The distributional statistics produced by distout and svydistout are the mean, P10, P20, P30, P50, P70 and P90 percentiles of the continuous outcome variable. Both cross-sectional and over time change analysis of the distribution is undertaken. The cross-sectional analysis produces statistics for the mean and various percentiles of the distribution of the continuous variable *varname* overall and for each of the groups specified by the variables included in the varlist for each year in the dataset (or for the years specified by the user), as well as differences in the each distributional statistics between the different subgroups defined by each variable included in the *varlist* (along with the corresponding standard errors and p-values). The over-time analysis produces statistics for the change in the mean and the various percentiles of the distribution of the continuous variable overall and for each of the groups defined by the variables included in the *varlist* for each year in the dataset (or for the years specified by the user) relative to the previous year, as well as statistics for the differences in the change of each distributional statistics of subgroups relative to the reference group of 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 version of the programme are available (distout and svdistout respectively). The survey design version of the programme (svydistout) is undertaking survey correction only for the mean and the change in the mean estimates.

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 distout and svydistout .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).

Syntax

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

Options

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

Important notes

svydistout:

Strata and psu variables have to be named strata and psu respectively.
svydistout is undertaking survey correction only for the mean and the change in the mean estimates (due to unavailability of svy version of the commands used to produce the remaining estimates).

distout and svydistout:

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.

Description

distout estimates and automatically saves in the current directory two excel files [named *varname*.xls] a range of distributional statistics for the continuous variable *varname* overall and for each of the groups specified by each variable of the *varlist* (typical variables of which can be included in the *varlist* include gender, age group, socio-economic status etc.). Specifying the bygroup (*groupvar*) option performs the distributional analysis separately for each subgroup in the *groupvar*. The distributional statistics produced by ***distout*** and ***svydistout*** are the mean, P10, P20, P30, P50, P70 and P90 percentiles of the *varname*. Both cross-sectional and over time analysis is undertaken. The estimates produced by the cross-sectional and over time analysis are described below.

Cross-sectional distributional analysis estimates: ***distout*** produces distributional statistics for the continuous variable *varname* for each of the group

identified by the *varlist* for each year included in the dataset - or for the years specified by the [**if** *exp*]. The estimates produced are the mean, the 10^{th}, 20^{th}, 30^{th}, 50^{th}, 70^{th} and the 90^{th} percentiles of the *varname*, along with their standard errors and their corresponding p-values.

Across groups differences relative to the reference category for each of the subpopulation identified by the categories in each variable in *varlist* are also produced along with their standard errors and the corresponding p-values. To estimate the across group differences in the mean we estimate the mean of the *varname* for each subpopulation identified by the categories in each variable in the *varlist* and test the across group diffidence in the estimates using the **lincom** stata command. To estimate the across the group differences in the percentiles we estimate a quantile regression using the qreg regression in stata with *varname* as a dependent variable and each variable in the *varlist* as independent variable. ***svydistout*** estimates average group differences in the means are performed by subpopulation analysis.

Over-time distributional analysis estimates: In addition to the cross-sectional estimates described above ***distout*** also produces estimates for the change in the mean and the various percentiles (10th, 20th, 30th, 50th, 70th and the 90th) of the continuous variable *varname* for each year t relative to year t-1 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 [**if** exp] option in order to examine patterns of change for years not covered by the default. The estimate of change in the mean of the *varname* for each subpopulation is undertaken using a mean estimation over year. The estimate of change in the percentiles of the *varname* is estimated using a quantile regression with the *varname* as a dependent variable and dummy variable which takes the value of 1 for year t and 0 for year t-1. In ***svydistout*** the estimates of the change in the mean are performed using subpopulation analysis.

Across groups differences in the change 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 estimate of difference change in the mean of the *varname* for each subpopulation relative to the base category for each variable in characteristic varlist is undertaken using an OLS regression model with the *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 mean and percentiles points estimates

Mean estimates:

mean *varname* if year==`y' [pw=`wgt'],over(`var', nolab)

svy: mean *varname* if year==`y',over(`var', nolab)

for each year `y' in the dataset and for each variable `var' in the *varlist*

Percentile point estimates:

qreg *varname* if `var'==`i' & year==`y' [pw=`wgt'],quantile(`q') vce(r)

Cross-sectional across group differences (relative to the base category of each variable in the varlist)

Tests for across group differences in the mean of the varname

distout

mean *varname* if year==`y' [pw=`wgt'],over(`var', nolab)

lincom [`outcome']`i'-[`outcome']`cmin'

svydistout

svy: mean `outcome' if year==`y',over(`var', nolab)

lincom [`outcome']`i'-[`outcome']`cmin'

for each year `y' in the dataset and for each variable `var' in the *varlist*

Testing percentile point differences

qreg *varname* i.`var' if year==`y' [pw=`wgt'],quantile(`q') vce(r)

Over time change estimates

Change in the mean

distout

xi: mean *varname* if `var'==`i' [pw=`wgt'],over(year)

lincom [`outcome']`y'-[`outcome']1

svydistout

subpop(subpop): mean `outcome',over(year)

where subpop is each group defined by each `var' in varlist

Change in the percentiles values

qreg *varname'* i.year if `var'==`i' [pw=`wgt'],quantile(`q') vce(r)

Over time change estimates

Cross-group difference in the change of the mean

reg *varname* i.year i.var i.year#i.var [pw=`wgt']

Cross-group difference in the change of the percentile

qreg varname i.year i.var i.year#i.var [pw=`wgt'],quantile(`q') vce(r)

Results saved in excel spreadsheets

Cross-sectional e*stimates:* The mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] are saved under the names: meanYr[year], P10Yr[year], P25Yr[year], P50Yr[year], P75Yr[year], P90Yr[year] respectively

***Standard errors of the cross-sectional estimates**:* The standard error of the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] are saved under the names: meanseYr[year], P10seYr[year], P25seYr[year], P50seYr[year], P75seYr[year], P90seYr[year]

***P-values of the cross-sectional estimates**:* The p-values of the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] are saved under the names: meanpvYr[year], P10pvYr[year], P25pvYr[year], P50pvYr[year], P75pvYr[year], P90pvYr[year]

Across group difference estimate: The difference in the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] for each group defined by each variable in the *varlist* relative to the reference category of each variable are saved under the names: dmeanYr[year], dP10Yr[year], dP25Yr[year], dP50Yr[year], dP75Yr[year], dP90Yr[year]

Standard errors of the across group difference estimate: The standard error of the difference in the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] for each group defined by each variable in the *varlist* relative to the reference category of each variable are saved under the names: dmeanseYr[year], dP10seYr[year], dP25seYr[year], dP50seYr[year], dP75seYr[year], dP90seYr[year]

***P-values of the across group difference estimate**:* The p-values of difference in the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] for each group defined by each variable in the *varlist* relative to the reference category of each variable in the varlist are saved under the names: dmeanpvYr[year], dP10pvYr[year], dP25pvYr[year], dP50pvYr[year], dP75pvYr[year], dP90pvYr[year]

Change in the estimate: The change in the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] relative to the previous year for each group defined by each variable in the *varlist* saved under the names:

cmeanYr[year], cP10Yr[year], cP25Yr[year], cP50Yr[year], cP75Yr[year], cP90Yr[year]

Standard errors of the change in the estimates: The standard errors of the change in the mean, P10, P25, P50, P75 and P90 percentiles of the varname for each year [year] relative to the previous year for each group defined by each variable in the varlist saved under the names: cmeanseYr[year], cP10seYr[year], cP25seYr[year], cP50seYr[year], cP75seYr[year], cP90seYr[year]

P-values of the change in the estimates: The p-value of the change in the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] relative to the previous year for each group defined by each variable in the *varlist* saved under the names: cmeanpvYr[year], cP10pvYr[year], cP25pvYr[year], cP50pvYr[year], cP75pvYr[year], cP90pvYr[year]

Across group difference in the change estimate: The difference in the change in the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] relative to the previous year for each group defined by each variable in the *varlist* relative to the reference category are saved under the names: cdmeanYr[year], cdP10Yr[year], cdP25Yr[year], cdP50Yr[year], cdP75Yr[year], cdP90Yr[year]

Standard errors of the across group difference estimate: The standard error of the difference in the change in the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] relative to the previous year for each group defined by each variable in the *varlist* relative to the reference category are saved under the names: cdmeanseYr[year], cdP10seYr[year], cdP25seYr[year], cdP50seYr[year], cdP75seYr[year], cdP90seYr[year]

P-values of the across group difference estimate: The p-values of the difference in the change in the mean, P10, P25, P50, P75 and P90 percentiles of the *varname* for each year [year] relative to the previous year for each group defined by each variable in the *varlist* relative to the reference category are saved under the names: cdmeanpvYr[year], cdP10pvYr[year], cdP25pvYr[year], cdP50pvYr[year], cdP75pvYr[year], cdP90pvYr[year]

Explanation of the excel files templates

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 the layout1: Cross-sectional results for each year [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 Yr[year] Change from base year.

Layout 2: Results are organised by distributional statistics

Name of the excel file: varname_Layout2.xls

Description of the layout2: Cross-sectional and over time estimates for each distributional statistic of the *varname* are saved in different excel worksheets under the name of each statistics: The worksheet named "Mean" holds all the results for the mean. Similarly the worksheets "P10", "P20", "P50", "P70" and "P90" hold results for the P10, P20, P50, P70 and P90 respectively.