

## Lab 5: Poverty and Inequality Analysis

Simon Halliday

ECO311, Fall 2016

### Revision Quiz

In your notes for your exercises for this lab, try to answer the following quiz questions based on Lab 4 *before class* and *without checking the lab!* In your exercises, have three categories for each question, "M" for my answer, "TPS" for think-pair-share, and "N" for after notes.

a. Explain what an SRS survey is with *epsem* sampling.

b. Explain in English what the following command does:

```
mean edyears [w = w1_wgt]
```

c. Explain in English what you think the following commands do :

```
svyset w1_hhcluster [pw = w1_wgt], strata(w1_hhdc)
egen hh_one = tag(hhid)
svy: mean plumbing_dum if hh_one == 1
```

d. Explain what you would do to complete the following tasks:

- i. find the mean of a per-capita income (pcy) using probability weights with the variable wt
- ii. understand the design effects after finding a mean of a variable
- iii. find a frequency table of access to plumbing (plumbing\_dum) using analytic weights with the variable wt
- iv. find a confidence interval of plumbing access after setting the cluster as the household cluster in the survey settings
- v. how would you find mean per-capita income for the sub-population women after setting up a survey design

When you get to class, check your answers with your neighbors (Think-Pair-Share). Once you have checked with your neighbors, go back to the lab to check what you didn't remember from the lab notes.

### Introduction

Before you start this lab session remember to do the following:

- Direct your working directory to the Lab5 folder (or Labs folder if you prefer)
- Make sure that your directory structure is consistent with Lab1 (Command-Files, log, cmdlog, etc) so you can use *relative file paths*.

### Doing Stata

By the end of this lab, you should:

- know about the different ways to measure inequality graphically and numerically
- understand the use of the DASP package to measure the Gini, Theil and Atkinson measures
- see the usefulness of a cumulative distribution function to understand poverty and income distribution
- identify the ways to measure poverty using the Foster-Greer-Thorbecke measures for the poverty headcount ratio, depth of poverty, and severity of poverty
- use quintiles to construct your own poverty lines

- Open a log file to track your work:

```
log using ./ProcessingAnalysis/CommandFiles/logs/
yourlastname_lab4.log
```

Be sure to specify the file extension .log

- Open a cmdlog file

```
cmdlog using ./ProcessingAnalysis/CommandFiles/cmdlogs/
yourlastname_lab4.do
```

After you have followed these steps, copy the NIDS data into your Original-Data folder for Lab4 and then open the data.

### Preparation

Run the following commands, which we'll need for variables we want to use in this lab.

```
gen death = w1_h_c1 == 1
replace death = . if w1_h_c1 == . | w1_h_c1 < 0
recode w1_r_b7 (-9/-3 24 =.) (25 = 0) (13 16 = 10) (14 17
= 11) (15 = 12) (18 = 13) (19 = 14) (20 = 15) (21 22
= 16) (23 = 17), gen(edyears)
gen electricity = w1_h_d22 == 1
replace electricity = . if w1_h_d22 == . | w1_h_d22 < 0
egen hh_one = tag(hhid)
gen pcminc = w1_hhincome / w1_hhsizer
```

We're going to need to use weights every so often, so let's generate weight variable that is the integer (int) of the variable w1\_wgt. You may recall that some of the weight commands need the weights to be integers in order to work properly.

```
gen weight = int(w1_wgt)
tab race [w = weight]
```

### The DASP Package

To analyze poverty and inequality, we shall use the DASP: Distributive Analysis Stata Package. DASP was developed by the World Bank and other organizations free of charge to help researchers and government workers to understand poverty and inequality within their countries. Poverty and inequality are part of the question about *distribution* of income (rather than the question of *efficiency*). DASP has to be installed manually and fortunately we have arranged to have it pre-installed on the machines we use in our class

REMINDER You should have learned in intermediate microeconomics about *Pareto efficiency* and other measures of efficiency (productive/allocative). An allocation is Pareto efficient if there exist no alternative allocation that is Pareto superior to it: an allocation is Pareto superior to another if at least one person can be made better off without making a person worse off. In economics, efficiency is often considered separate from questions of *distribution*. Often, the question becomes: can efficiency and distribution be improved simultaneously, or are there trade-offs between alleviating poverty and improving efficiency?

venues. Depending on your research question, you may also choose to use the `ineqdeco` (inequality decomposition) package to understand issues of distribution.

### *Inequality*

There are many resources online to investigate inequality. The Food and Agricultural Organization (FAO) has many resources I would recommend you look at:

- Basics of inequality measurement: [www.fao.org/docs/up/easypol/448/simple\\_inequality\\_mesures\\_080en.pdf](http://www.fao.org/docs/up/easypol/448/simple_inequality_mesures_080en.pdf)
- Inequality and axioms for its measurement: [www.fao.org/docs/up/easypol/447/inqulty\\_axms\\_msrmt\\_054en.pdf](http://www.fao.org/docs/up/easypol/447/inqulty_axms_msrmt_054en.pdf)
- Theil Index and Entropy Class measures of inequality: [www.fao.org/docs/up/easypol/445/theil\\_index\\_051en.pdf](http://www.fao.org/docs/up/easypol/445/theil_index_051en.pdf)
- Welfare measures of inequality: The Atkinson measure [www.fao.org/docs/up/easypol/451/welfare\\_measures\\_inequa\\_atkinson\\_050en.pdf](http://www.fao.org/docs/up/easypol/451/welfare_measures_inequa_atkinson_050en.pdf)
- Decomposition of income inequality by subgroups: [http://www.fao.org/docs/up/easypol/444/dcmpsng-inqulty\\_sbggrp\\_052en.pdf](http://www.fao.org/docs/up/easypol/444/dcmpsng-inqulty_sbggrp_052en.pdf)

### *Lorenz Curves*

Now we can discuss the example of an inequality analysis using per capita total monthly income as our choice of measure of well-being. See the DASP User Manual on Moodle.

First, in case you've forgotten how to set up a survey, you can ask for help: `help svyset`. Otherwise, we shall set up the survey as follows:

```
svyset w1_hhcluster [pweight = weight]
```

Inspect the income variable we generated using `sum` or `tabsum` with a cross-tab.

Now, we can use DASP to produce a kernel density plot:

```
cdensity lnpcmnc , type(den) min(0) max(12)
```

#### **Exercise 1**

1. What does a cumulative distribution function do?
2. What does the cumulative distribution function tell us about the distribution of income?

Having looked at cumulative distribution functions, we can try to map the cumulative population (x-axis) against the cumulative share of income (y-axis). This allows us to graph a *Lorenz* curve. Having graphed a Lorenz curve we will be able (later) to calculate the Gini index.

DASP makes it easy to plot Lorenz Curves while recognizing the survey weights (this would be pretty hard to do by hand).

```
clorenzs pcminc
```

What we have done here is asked DASP (and Stata) to produce a Lorenz curve using the variable per capita monthly income (pcminc). We can actually do this for any continuous variable, which researchers have also done for wealth and for years of education.

Analyzing inequality and looking at Lorenz curves also allows us to compare visually whether there are differences in inequality across groups. For example, what does the Lorenz curve of white South Africans look like versus the Lorenz curve for Black South Africans? When we compare Lorenz curves (and inequality) in this way, we are said to be looking for 'inequality dominance.' That is, is the level of inequality better or worse for one group rather than another, or does the inequality in one group *dominate* the inequality in another group?

To analyze inequality dominance by race we need to plot Lorenz curves for all race groups. We can do this with the group variable equal to race as follows:

```
clorenz pcminc , hgroup(race)
```

You can check the weights by checking the svyset option.

### Exercise 2

Is there evidence of inequality dominance by race? Explain what you see and how you would motivate your position.

### *Measuring Inequality*

We can go from the Lorenz curve to inequality measures that condense the income distribution into one statistic. We can look at a variety of different measures, many of which are explained in the manual, go to: DASP/inequality/Gini/generalisedentropy/Atkinson

Consider what happens when you type the following different commands:

```
igini pcminc
ientropy pcminc , theta(1)
iatkinson pcminc , epsilon(1)
```

What do you see?

As with the Lorenz curves, we may want to break down the inequality measurements by groups, such as race/ethnicity or gender.

```

igini pcminc, hgroup(race)
ientropy pcminc, theta(1) hgroup(race)
iatkinson pcminc, epsilon(1) hgroup(race)

```

### *Between and Within Group decompositions*

We can break down inequality both *between* groups and *within* groups.

For example, consider the race groups in South Africa. Among white people in South Africa, some people have higher income than others, so there will be inequality *within* the white group. Similarly, for black Africans, there will be income inequality *within* the group. But, given that we have seen that the level of inequality in the country as a whole and the levels of inequality within the two groups differ, what explains these differences? We need to understand *between* group inequality, that is, comparing white and black South Africans, how much of the inequality in income differs *across* or *between* the two groups?

When we compiled Lorenz curves for each race, we could use that to find a Gini index for each race, but this did not tell us whether there's more inequality between black and white people or within each group.

To do within and between group decomposition, we need to use the Theil measures of inequality. We can do this using the DASP package, which you can look up in the DASP manual at `dasp/decomposition/generalisedentropy` with groups and setting two levels of theta: `theta = 0` and `theta = 1`, for example:

```

dentropyg pcminc, hgroup(race) theta(0)
dentropyg pcminc, hgroup(race) theta(1)

```

Compare these two in detail.

To see how the decomposition is done, let us take the mean log deviation (a.k.a.  $GE(0)$ ) which is the easiest to decompose. To calculate the within-group inequality, we sum the inequality measure for each group, weighted by the population share:  $GE(0)_{within} = (PopulationshareofAfricans) * (GE(0)_{Africans}) + (Pop.ShareofColoured) * (GE(0)_{Coloured}) + \dots$  Plugging in the numbers from the attached output, we get:  $GE(0)_{within} = ?$

Looking at the "Within-group" inequality table under  $GE(0)$ , we see that STATA concurs with our calculations. To get the between-group inequality, we simply subtract the within-group inequality from the total inequality.  $GE(0)_{between} = GE(0)_{total} - GE(0)_{within} = 0.9350 - .6197 = .3152$ . We find the same number in the "Between-group" table under  $GE(0)$ . So, what does this tell us? Well, now we can say that about 66% ( $= .6197/.9350$ ) of inequality in the distribution of per-capita income (as measured by the mean log deviation) is within-race inequality. The other 34% is between-race. We could not have said that just by looking at Lorenz curves.

There is no Atkinson decomposition. So, if you want to use the Atkinson

See `DASP/decomposition/generalisedentropydecompositionbygroups`.

The Theil inequality measures fall within a broader class of inequality measures called the *generalized entropy measures* of inequality, hence the use of the word 'entropy' in the Stata command. The

measure, you need to use another package such as `ineqdeco`.

## Poverty

As with inequality, the FAO has very useful handbooks for understanding poverty:

- Basic Poverty Measures: [www.fao.org/docs/up/easypol/323/basic-pov-measures\\_007EN.pdf](http://www.fao.org/docs/up/easypol/323/basic-pov-measures_007EN.pdf)
- Generalized Poverty gap measures: [www.fao.org/docs/up/easypol/324/gnrised\\_pvty\\_gap\\_msr\\_010EN.pdf](http://www.fao.org/docs/up/easypol/324/gnrised_pvty_gap_msr_010EN.pdf)
- Poverty and dominance: [www.fao.org/docs/up/easypol/431/povetyanddominance\\_035en.pdf](http://www.fao.org/docs/up/easypol/431/povetyanddominance_035en.pdf)

First of all, we're going to generate a new variable, the natural log of income `lnpcminc`, the natural log of per capita income:

```
gen lnpcminc = log (pcminc)
```

As we did with `pcminc` in levels previously, let's look at a cumulative distribution function of the log of income:

```
cdensity lnpcminc , type(den) min(0) max(12)
```

The cumulative distribution function (CDF), that is, the proportion of the sample that have a value of the variable less than or equal to each value that the variable takes on, is the starting point of poverty analysis and the basis of *poverty dominance* analysis.

Now we shall use the DASP package and combine it with poverty measurements. We're going to use per capita monthly income and set a maximum level for the graphic:

Now let us generate a CDF for the `pcminc` variable

- Consult: `dasp/curves/fgt`
- set the maximum at ZAR2000 (2000 South African Rands – the local currency – equivalent to roughly \$200 per month at the time)

We can do this with the following:

```
cfgt pcminc , alpha(0) min(0) max(2000)
```

The command above uses the 'FGT' measures, where FGT stands for Foster-Greer-Thorbecke. The Foster-Greer-Thorbecke measure is a measure of poverty that takes different values depending on a value of alpha ( $\alpha$ ) in  $FGT(\alpha)$ . The equation for the FGT measure is as follows:

$$FGT(\alpha) = \frac{1}{Nz^\alpha} \sum_{i=1}^q (z - y_i)^\alpha \quad (1)$$

For the equation,  $N$  is the total population,  $z$  is the poverty line,  $y_i$  is the income level of person  $i$  and  $q$  is the total number of people who are poor (have incomes below the poverty line,  $z$ ).

There are three levels of alpha that we care about:

$\alpha = 0$  is the *poverty headcount ratio* (FGT(0)), which means you want to understand the proportion of people who are poor within the population, that is, you simply count all individuals with income below the poverty line and divide that total by the total population.

$\alpha = 1$  is the *poverty gap index* (FGT(1)), which measures the *depth* of poverty. It computes for each individual how far they are from the poverty line. It sums up all of these individual poverty gaps to find the total poverty gap. It then divides the total poverty gap by the number of people in the country multiplied by the poverty line to find a number between 0 and 1 (the closer to 1 the number is, the greater the depth of poverty).

$\alpha = 2$  is the *poverty gap-squared index* (FGT(2)), which measures the *severity* of poverty. It computes each individual's poverty gap (the distance between their income and the poverty line) and squares that number. This results in greater weight being given to people whose incomes are further from the poverty line (that is, to poorer individuals). The squared poverty gaps of all individuals are summed and divided by the total population multiplied by the poverty line squared. This results in a number between 0 and 1, which suggests that poverty is more severe the closer the number is to 1.

Now let's explore poverty dominance

- `dasp/curves/fgts`
- You can look at differences between CDFs using the `hgroup` option, which constructs separate CDFs for the different groups that you specify. The groups we shall specify are South Africa's racial groups.

Type the following:

```
group variable race
cfgt pccinc, alpha(0) hgroup(race) min(0) max(2000)
```

### Exercise 3

Examining the different CDFs, what can you say about the different incomes of the different "racial groups" in South Africa.

You can place 'poverty lines' on the graph using the `xline` option:

```
cfgt pccinc, alpha(0) hgroup(race) min(0) max(2000) xline
(515 949)
```

The `xline` command simply adds a vertical line at that value on the x-axis, which serves the purpose that a poverty line would (it tells us what proportion of the population is above or below the line). Notice that we use two different lines, a lower one (ZAR515) and an upper one (ZAR949). These are the lines stipulated by the South African government, corresponding roughly to the World Bank poverty lines for poverty (food, clothing and shelter) and extreme poverty (food and clothing only).

#### Exercise 4

- a. At the two poverty lines, what appears to be true about the different experiences of poverty across the different racial groups?
- b. Comparing the CDFs by group and the CDF without groups, which group do you think is 'driving' the poverty in South Africa? What approximate proportion of this group appear to be poor at the two poverty lines? Compare their proportion of poor to those in other groups before looking again at the CDF without groups.

Now that we have the results that we want, we can make sure that the graph looks good by adding titles, a legend, and other details (go back to your previous labs to check out how to do this – we covered it in Lab 2).

### *Measuring Poverty*

Cumulative distributions are extremely useful for getting a feel for the whole distribution of income (or consumption, or whatever else you might be interested in). But there are also advantages to condensing this information into a single statistic such as a poverty rate, especially if the CDFs for two groups cross each other. One advantage is that it is easier to summarize what is going on with a single number: you can do more acrobatics with them. For example, we can say how much poverty is accounted for by different subgroups of the population once we choose to define poverty in a particular way. In this section, we are going to get Stata to calculate different poverty measures, and then see what we can do with them.

Earlier, we informally condensed cumulative distributions into a single poverty measure by picking a poverty line and reading off the proportion that earned less than that amount from the graph. That proportion corresponds to one particular poverty measure: the headcount ratio.

We can calculate this and several other poverty measures by using DASP

- `DASP/poverty/FGT`
- You have to specify group variable, such as race, an alpha (whether you want to measure the headcount ratio, poverty gap index, or poverty gap squared index) and a poverty line (`pline = [a number]`)

Here's an example:



```
ifgt pcminc, alpha(0) hgroup(race) pline(515)
```

Compare poverty at ZAR515 with poverty at ZAR949:

```
ifgt pcminc, alpha(0) hgroup(race) pline(949)
```

Now we go on to calculate poverty decompositions and shares DASP/ decomposition/FGT decomposition by groups with different levels of alpha - the poverty headcount index, poverty gap index, and poverty gap-squared index. We do this by specifying different levels of alpha for the same poverty line (the same pline):

```
dfgtg pcminc, hgroup(race) alpha(0) pline(515)
```

Compare shares with alpha = 0 and alpha = 1:

```
dfgtg pcminc, hgroup(race) alpha(1) pline(515)
```

#### Exercise 5

Compare the depth of poverty (alpha = 1) and the severity of poverty (alpha = 2) for the two different poverty lines. What do you find?

#### *Defining your own poverty line*

First we sort the data using per capita monthly income (pcminc):

```
sort pcminc
```

Now we use the `xtile` command to generate quintiles of `pcminc`. A quintile is a 5th or slices of the distribution into 20 percentage point slices. So you have 0 – 20, 21 – 40, 41 – 60, 61 – 80, 81 – 100 for the rankings of per capita monthly income.

```
xtile quint = pcminc [w = weight], nq(5)
```

We can quickly examine the quintiles and see what the average per capita monthly incomes are in each:

```
tab quint [w = weight], sum(pcminc)
```

#### Exercise 6

What is the mean per capita monthly income in the lowest quintile?

What is the mean per capita monthly income in the second lowest quintile?

Having the mean doesn't necessarily tell us what we want for a poverty line, though. Often, we might want to use the maximum income at the top of a quintile range to use for a poverty line. We can find these out with the following commands, stipulating that we only want to look at the summary statistics if they are in the lowest quintile:

```
sum pcminc if quint==1
```

Our ultra-poor poverty line is ZAR238 per capita per month

Or we could look at the second lowest quintile.

```
sum pcminc if quint==2
```

Our overall poverty line is ZAR418 per capita per month

#### Exercise 7

What happens to the poverty measures you examined earlier if you use the two poverty lines we just found rather than poverty lines of ZAR515 and ZAR945?

#### Commands and options in Lab 4

weight (w, aw, pw, iw)	predict	twoway lfitci
xi: reg	char ... [omit]	ci
sys dir	net search	outreg2 using filename.extension
bdec(3)	excel	replace
append		label
estout	estimates store	putexcel
twoway, legend label		