Lab 4: Stochastic Dominance Testing
Short Course on Poverty & Development for Nordic Ph.D. Students
University of Copenhagen
June 13-23, 2000

In this lab we'll continue to use the same data that we used in the previous labs (e.g. `c:\copen\data\expend.dta`). In addition, we'll take advantage of the `c:\ado\fgt.ado` program that we wrote in Lab 3, and imbed it in a progam that we'll write to conduct stochastic dominance tests for poverty between independent distributions of expenditures.

Stata Notes: Not much to add here, except this will give us an example of how we can take printed output in a '.log' file and graph it in a spreadsheet application like Excel.

Approach: Here's a quick review of the concept of stochastic dominance testing, and how we'll apply it.

The idea of standard tests of welfare dominance to compare distributions of welfare indicators (inter-temporally or spatially) is to make ordinal judgments on how poverty changes for a wide class of poverty measures over a range of poverty lines. We start by discussing the concept of welfare dominance, and then explain how to estimate the orderings and to perform statistical inference on them. The discussion follows Ravallion (1994) and Davidson and Duclos (1998) closely.

Consider two distributions of welfare indicators with cumulative distribution functions, F_A and F_B, with support in the nonnegative real numbers. Let

$$D_A^1(x) = F_A(x) = \int_0^x dF_A(y).$$

If $D_A^1(x) \le (<) D_B^1(x)$ for all $x \in \Re_+$ (i.e. F_A is everywhere to the right of F_B), then distribution *A* is said to (strictly) first order dominate distribution *B*. In terms of welfare economics, the interpretation is that up to the poverty line *x*, *A* is a better distribution than *B* for any welfare function that is both increasing in the welfare variable (e.g. expenditures) and anonymous, in the sense that we do not care that one particular person's welfare falls, as long as another's rises by more than enough to compensate. If we can say this for a broad range of poverty lines, then we have a quite general conclusion that *A* is preferable to *B*.

Since $D_A^1(x)$ is also the poverty *headcount* ratio (P_0) where the *x* is the poverty line, it follows that first order dominance implies that poverty as measured by P_0 is lower for distribution *A* than for distribution *B* regardless of the poverty line chosen. Dominance results can also be considered up to a maximum allowable poverty line if we aren't concerned with relative changes in the upper ends of the distribution.

If the two distributions cross within the range of poverty lines that we consider relevant, then first order dominance does not hold, and we know that different poverty lines and measures will rank the distributions differently. In other words, depending on the poverty line or measure chosen, we might simultaneously conclude that poverty increased or decreased. In this case, we can still make a fairly general welfare statement if second order dominance holds. In particular, if *A* second-order dominates *B*, then *A* is a better distribution than *B* for all welfare functions that are increasing, anonymous, and that favor equality. To define second-order dominance, let $D_A^2(x)$ be the area under F_A up to *x*,

$$D_A^2(x) = \int_0^x D_A^1(y)dy .$$

If $D_A^2(x) \le (<) D_B^2(x)$ for all *x* (i.e. the area under F_A up to *x* is less the area under F_B up to *x*), then distribution *A* is said to (strictly) second order dominate distribution *B*.

If, to use Ravallion's (1994) terminology, the "poverty deficit" curves (D^2) cross, then higher orders of dominance can be checked. To generalize, let

$$D_A^s(x) = \int_0^x D_A^{s-1}(y)dy ,$$

for any integer, $s \ge 2$. Now distribution *A* is said to (strictly) dominate distribution *B* at order *s* if $D_A^s(x) \le (<) D_B^s(x)$.

Davidson and Duclos (1998) show that $D^s(x)$ can be equivalently expressed as

$$D^s(x) = \tfrac{1}{(s-1)!}\int_0^x (x-y)^{s-1} dF(y) .$$

This formulation makes it easy to see that second order dominance implies that the *poverty gap* (P_1) is less for distribution *A* than for distribution *B* for all possible poverty lines. Further, third order dominance implies an unambiguous change in the *squared poverty gap* (P_2). To generalize even further, and as we showed in the last lab, welfare dominance of order *s* implies that the Foster-Greer-Thorbecke poverty measure P_{s-1} is less for distribution *A* than for distribution *B* for all possible poverty lines. Foster and Shorrocks (1988) show that while first-order dominance is a sufficient condition for higher-order dominance, it is not a necessary condition. Thus if we find that a distribution first-order dominates another, then we know how poverty as measured by any of the FGT P_a measures has changed over the relevant range of poverty lines.

Davidson and Duclos (1998) also show that if we have a random sample of *N* independent observations on the welfare variable, y_i, from a population, then a natural estimator of $D^s(x)$ is

$$\hat{D}^s(x) = \frac{1}{N(s-1)!}\int_0^x (x-y)^{s-1}\,d\hat{F}(y)$$

$$= \frac{1}{N(s-1)!}\sum_{i=1}^{N}(x-y_i)^{s-1} I(y_i \le x)$$

where $\hat{F}$ is the empirical cumulative distribution function of the sample, and $I(\cdot)$ is an indicator function, which is equal to one when it's argument is true, and equal to zero when false.

We apply this estimator to two independent samples for each of our indicators. Thus the variance of the difference of the two estimators is,

$$\text{var}(\hat{D}^s_A(x) - \hat{D}^s_B(x)) = \text{var}(\hat{D}^s_A(x)) + \text{var}(\hat{D}^s_B(x)),$$

which is easy to estimate since $\hat{D}^s(x)$ is a sum of *iid* variables (as we showed in Lab 3). Simple *t* statistics are constructed to test the null hypothesis,

$$H_0 : \hat{D}^s_A(x) - \hat{D}^s_B(x) = 0,$$

for a series of test points up to an arbitrarily defined highest reasonable poverty line. In cases where the null hypothesis is rejected and the signs are the same on all of the *t* statistics, then dominance of order *s* is declared.

Now we'll write a program do test for stochastic dominance up to order 3:

```
#delimit ;
program define dom;   * syntax is dom var1 wgt1 var2 wgt2 zmax #test-points;
  version 6.0;
  local x1     `1';
  local wgt1   `2';
  local x2     `3';
  local wgt2   `4';
  local zmax = `5';
  local num  = `6';
 summarize `x1', meanonly;
 local min1   = r(min);
 summarize `x2', meanonly;
 local min2   = r(min);
 local min    = max(`min1',`min2');
 local inter  = (`zmax'-`min')/(`num'-1);
 local min    = `min’ + `inter’;         * Don’t start at very end of tail;
 local inter  = (`zmax'-`min')/(`num'-1);
 local done   = 0;
 local s      = 1;
 tempvar ind;

display " ";
display "Minimum test point is " `min';
display " ";
display "Maximum test point is " `zmax';
display " ";
```

```
 if `zmax' <= `min' {;
        local done = 1;
        display " ";
        display in red "Error: " in yellow "Max test point is not larger than
min test point!";
        display " ";
              };

 while `done' == 0 {;
        display " ";
        display "Order " `s';
        display "Z          D1        D2        t-statistic";
        local alpha = `s' - 1;
        local  i    = 1;
        local  z    = `min';
        quietly gen `ind' = .;
        while `z'  <= `zmax' {;
              quietly fgt `x1' `wgt1' `z' `alpha';
              local  D1 =  ${D`s'};
              local VD1 =  ${VD`s'};
              quietly fgt `x2' `wgt2' `z' `alpha';
              local  D2 =  ${D`s'};
              local VD2 =  ${VD`s'};
              local   t = (`D2'-`D1')/sqrt(`VD1'+`VD2');
              display %5.4f `z' "    " %5.4f `D1' "    " %5.4f `D2'  "    "
%4.2f `t';
              local ind0 = 1.5;
              if `t' >=  1.96 {; local ind0 = 1; };
              if `t' <= -1.96 {; local ind0 = 2; };
              quietly replace `ind' = `ind0' if _n==`i';
              local z = `z' + `inter';
              local i = `i' + 1;
              };
        summarize `ind', meanonly;
         if r(mean) == 1 {; local done = 1;
                             global Dorder = -`s';
                             display "";
                             display "Dominance achieved at order " `s';
                             local s = 1;};
         if r(mean) == 2 {; local done = 1;
                             global Dorder = `s';
                             display "";
                             display "Dominance achieved at order " `s';
                             local s = 1;};
         if `s' >= 3        {; local done = 1;
                             global Dorder = 0;
                             display "";
                             display "Dominance not achieved up to order
3";};
        drop `ind';
        local s = `s' + 1;
        };
end;
```

Now let's use this program to test the difference between the distributions for regions 1 and 2:

```
use c:\copen\data\expend
gen pcexp = hhexpend / hhsize
sum pcexp, detail
scalar zmax = r(p50)
preserve
keep if region==1
keep pcexp hhsize
rename pcexp pcexp1
rename hhsize hhsize1
save c:\temp\temp, replace
restore
preserve
keep if region==2
keep pcexp hhsize
merge using c:\temp\temp
dom pcexp1 hhsize1 pcexp hhsize zmax 40
restore
```

The output should look like

```
. dom pcexp1 hhsize1 pcexp hhsize zmax 40

Minimum test point is 280.77197

Maximum test point is 1144.3576


Order 1
Z           D1        D2        t-statistic
302.9152    0.0107    0.0006    -2.63
325.0584    0.0169    0.0016    -3.14
347.2016    0.0196    0.0016    -3.45
369.3449    0.0271    0.0016    -4.19
391.4881    0.0338    0.0016    -4.77
413.6313    0.0440    0.0016    -5.55
435.7745    0.0532    0.0039    -5.82
457.9177    0.0655    0.0064    -6.28
480.0610    0.0714    0.0072    -6.55
502.2042    0.0838    0.0148    -6.40
524.3474    0.0977    0.0206    -6.60
546.4906    0.1111    0.0305    -6.40
568.6339    0.1323    0.0432    -6.47
590.7771    0.1444    0.0519    -6.41
612.9203    0.1699    0.0677    -6.55
635.0635    0.1844    0.0844    -6.11
657.2067    0.2129    0.0977    -6.65
679.3500    0.2491    0.1119    -7.49
701.4932    0.2770    0.1298    -7.71
723.6364    0.3063    0.1564    -7.52
745.7796    0.3211    0.1786    -6.98
767.9228    0.3417    0.1994    -6.81
790.0661    0.3774    0.2268    -7.00
812.2093    0.4169    0.2506    -7.55
834.3525    0.4360    0.2737    -7.27
856.4957    0.4566    0.2998    -6.95
878.6389    0.4803    0.3161    -7.22
900.7822    0.5114    0.3304    -7.93
922.9254    0.5291    0.3399    -8.27
```

```
945.0686    0.5554    0.3628    -8.40
967.2118    0.5745    0.3813    -8.42
989.3551    0.5922    0.3992    -8.42
1011.4983    0.6059    0.4181    -8.20
1033.6415    0.6177    0.4360    -7.95
1055.7847    0.6360    0.4588    -7.78
1077.9279    0.6542    0.4681    -8.23
1100.0712    0.6703    0.4805    -8.44
1122.2144    0.6875    0.4932    -8.71

Dominance achieved at order 1
```

The way to interpret this is that distribution 1 (region 1) has more poverty than distribution 2 (region 2) for all poverty lines up to 1,122 and for all poverty measures. If the differences at certain test points were not statistically significant, we couldn't say for sure that the curves didn't cross there and we couldn't declare dominance. In this case, the differences between the distributions are statistically significant for all test points.

Note that the output for the "Order 1" test gives us the estimated CDFs. These can be plotted in Excel to give an idea of what's going on in the tests. There are two ways to plot these CDFs: (1) Cut and past the output into a separate text (.txt) file and open this using Excel, and (2) Highlight the output in the "Stata Results" window, copy it by going to `Edit`, then `Copy Table` (or `CTRL+SHIFT+C` on the keyboard), and paste it in an Excel worksheet. Then you can use Excel commands to graph `D1` and `D2` on `Z`.

Now experiment with the remaining regions and urban-rural.