Group Comparisons:
Differences in Composition Versus Differences in Models and Effects

Overview. This is the first of a series of handouts that will deal with techniques for comparing groups. This initial handout notes that, when comparing groups, it is important to realize that groups can differ in two ways:

- There can be compositional differences between groups. Specifically, means on the IVs may differ between groups.
- The effects of IVs can differ between groups. A variable might have a stronger effect on one group than it does on the other. Indeed, the direction of an effect may even differ between groups. The model that describes one group may be very different from the model that describes another.

For example, blacks may have lower levels of education and less job experience than do whites. As a result, they may tend to have lower levels of income, even if the effects of education and job experience are the same for both groups. Simple T-tests or ANOVA tests can determine whether there are significant compositional differences between groups.

Or, blacks may have similar levels of education and job experience, but the effects of these variables may be less for them, e.g. a year of education is worth less to a black than it is to a white. As a result, blacks may tend to have lower incomes than comparable whites.

Compositional, or mean, differences between groups on the IVs may suggest that differences on the DVs are "justified", e.g. blacks earn less than whites because they are less educated; women earn less than men because they are concentrated in lower-paying occupations, or have less continuous service with the same company. Of course, one must then ask what produced the compositional differences.

Differences in effects raise questions about why those differences exist. If blacks benefit less from education than whites, is this perhaps because of discrimination? Or do other factors need to be considered in the model?

It is important to keep compositional differences and differences in effects separate. Researchers will sometimes confuse the two, muddling the discussion of why group differences exist. In particular, researchers sometimes focus a lot on their models, and overlook how important compositional factors can be in explaining group differences.

Differences in Composition. Returning again to our hypothetical data from 400 whites and 100 blacks: the following t-tests and descriptive statistics reveal that blacks have lower levels of education and job experience than do whites. These differences are all highly significant. These lower levels of education and job experience probably are part of the reason that black income is also lower than white income.

```
. use http://www.nd.edu/~rwilliam/stats2/statafiles/blwh.dta, clear

. ttest   educ, by(black)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   white |     400        13.9     .175505      3.5101    13.55497    14.24503
   black |     100        10.2    .4376244    4.376244    9.331658    11.06834
---------+--------------------------------------------------------------------
combined |     500       13.16     .178023    3.980715    12.81023    13.50977
---------+--------------------------------------------------------------------
    diff |                 3.7     .413502                2.887576    4.512424
------------------------------------------------------------------------------
Degrees of freedom: 498

                  Ho: mean(white) - mean(black) = diff = 0

     Ha: diff < 0               Ha: diff != 0              Ha: diff > 0
       t =   8.9480            t =   8.9480              t =   8.9480
   P < t =   1.0000          P > |t| =   0.0000          P > t =   0.0000

. ttest   jobexp, by(black)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   white |     400        14.1    .2395171    4.790341    13.62913    14.57087
   black |     100        11.2    .5464301    5.464301    10.11576    12.28424
---------+--------------------------------------------------------------------
combined |     500       13.52    .2263663    5.061703    13.07525    13.96475
---------+--------------------------------------------------------------------
    diff |                 2.9    .5513765                1.816689    3.983311
------------------------------------------------------------------------------
Degrees of freedom: 498

                  Ho: mean(white) - mean(black) = diff = 0

     Ha: diff < 0               Ha: diff != 0              Ha: diff > 0
       t =   5.2596            t =   5.2596              t =   5.2596
   P < t =   1.0000          P > |t| =   0.0000          P > t =   0.0000

. ttest   income, by(black)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   white |     400       30.04    .3897187    7.794375    29.27384    30.80616
   black |     100       18.79    .7664749    7.664749    17.26915    20.31085
---------+--------------------------------------------------------------------
combined |     500       27.79    .4013067    8.973491    27.00154    28.57846
---------+--------------------------------------------------------------------
    diff |               11.25    .8685758                9.543475    12.95652
------------------------------------------------------------------------------
Degrees of freedom: 498

                  Ho: mean(white) - mean(black) = diff = 0

     Ha: diff < 0               Ha: diff != 0              Ha: diff > 0
       t =  12.9522            t =  12.9522              t =  12.9522
   P < t =   1.0000          P > |t| =   0.0000          P > t =   0.0000
```

Differences in Effects. When doing group comparisons, we also often wish to test whether model parameters are the same in two or more populations. For example, we might want to test whether the effects of education and job experience on income are the same for both whites and blacks, i.e. do blacks get as much benefit from their education and job experience as do whites? There are various ways to do this. One procedure that will let us test whether there are <u>any</u> differences in effects across groups is as follows:

- Estimate separate regressions for blacks and whites.
    - Add up the error sums of squares from both groups; this is SSE_u.
    - Also note the sample size for each group, i.e. N_1 and N_2.
    - We refer to this as the unconstrained model, because coefficients are free to differ between populations.
    - For group 1, error d.f. = N_1 - K – 1
    - for group 2 error d.f. = N_2 - K – 1
    - hence the error d.f. for both together is $N_1 + N_2$ - 2K - 2.
    - Put another way, a total of 2K + 2 coefficients are estimated: 2 sets of betas, and 2 intercepts, hence the error d.f. in the unconstrained model = $N_1 + N_2$ - 2K - 2.
    - Put another way – we have Group 0 and Group 1. The unconstrained model is obtained by estimating the regressions

    $$Y = \alpha^{(0)} + \beta_1^{(0)} X_1 + \beta_2^{(0)} X_2 + \varepsilon \text{ for group 0}$$

    $$Y = \alpha^{(1)} + \beta_1^{(1)} X_1 + \beta_2^{(1)} X_2 + \varepsilon \text{ for group 1}$$

    where the superscripts stand for the group number.

- Estimate a regression for both groups together. This will give you SSE_c. We refer to this as the constrained model, because parameters (including the intercept) are constrained to be equal in both populations.
- Note that J (the number of restrictions) = K + 1. This is because, not only are all the X's constrained to have equal effects across groups, the intercepts are also constrained to be equal. Also, Total N = $N_1 + N_2$.
- You then compute the incremental F:

$$F_{K+1, N_1+N_2-2K-2} = \frac{(SSE_c - SSE_u) * (N_1 + N_2 - 2K - 2)}{SSE_u * (K+1)}$$

- If the F value is significant, you reject the null, and conclude that coefficients are not the same across groups.
- This strategy can easily be modified for more than 2 groups. Just run separate regressions for each group, add up the SSE's to get the unconstrained SSE. Remember that J = (Number of groups - 1) * (K + 1), unconstrained error d.f. = total sample size - [number of groups*(K + 1)].
- Note that the above (as well as approaches that will produce equivalent results) is sometimes called a *Chow Test*. A Chow Test is a test of whether any parameters differ across populations. I've also seen the term Chow Test used when the intercept is allowed to differ

across populations but all other parameters are constrained to be the same. To let the intercepts differ, in the constrained model you would add dummy variables for group membership.

There are two main concerns with this approach:

- The incremental F test does not tell you which coefficients differ across populations. Indeed, it does not even tell you whether it is one of the IV effects that differs across populations. It could just be that the intercept term differs in the two groups.
- You may think that some IVs have different effects in different groups, while other IVs have the same effects in each group. This approach does not allow you to test whether a subset of the variables has different effects across groups. For good theoretical reasons, you may believe that some effects will differ across groups, while others will not.

An alternative approach, which we will describe shortly, makes it possible to overcome these limitations. This approach uses interaction effects rather than estimating separate models for each group.

EXAMPLE. In our modified Income/Job experience/Education example, there are 100 blacks and 400 whites. First, we estimate separate regressions for blacks and whites. This is easily done in either Stata or SPSS. Using Stata,

```
. bysort black: regress income educ jobexp

--------------------------------------------------------------------------------------------------
-> black = white

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  2,   397) =  620.07
       Model |  18361.9894     2  9180.99472           Prob > F      =  0.0000
    Residual |  5878.16991   397  14.8064733           R-squared     =  0.7575
-------------+------------------------------           Adj R-squared =  0.7563
       Total |  24240.1594   399  60.7522791           Root MSE      =  3.8479

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.893338   .0562591    33.65   0.000     1.782735    2.003941
      jobexp |    .722255   .0412236    17.52   0.000     .6412111    .8032988
       _cons |  -6.461189   1.089219    -5.93   0.000    -8.602546   -4.319831
------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------
-> black = black

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =  267.80
       Model |  4924.27286     2  2462.13643           Prob > F      =  0.0000
    Residual |   891.81705    97  9.19399021           R-squared     =  0.8467
-------------+------------------------------           Adj R-squared =  0.8435
       Total |  5816.08991    99   58.748383           Root MSE      =  3.0322

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.677949   .0725479    23.13   0.000     1.533962    1.821936
      jobexp |    .421975   .0581021     7.26   0.000     .3066585    .5372915
       _cons |    -3.0512   1.154604    -2.64   0.010    -5.342771   -.7596303
------------------------------------------------------------------------------
```

Or alternatively, you could do

```
. regress income educ jobexp if black
. regress income educ jobexp if !black
```

Hence, for whites, N_w = 400, SSE_w = 5878.17, DFE_w = 397 . For blacks, N_b = 100, SSE_b = 891.82, DFE_b = 97. Combining the black and white numbers for the unconstrained model,

$$N_u = 500, SSE_u = 6770, DFE_u = 494.$$

For the constrained model, Income is regressed on Educ and Jobexp for both groups together:

```
. reg  income educ jobexp

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) = 1103.96
       Model |  32798.4018     2  16399.2009           Prob > F      =  0.0000
    Residual |  7382.84742   497  14.8548238           R-squared     =  0.8163
-------------+------------------------------           Adj R-squared =  0.8155
       Total |  40181.2493   499  80.5235456           Root MSE      =  3.8542

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    1.94512   .0436998    44.51   0.000     1.859261     2.03098
      jobexp |   .7082212   .0343672    20.61   0.000     .6406983     .775744
       _cons |  -7.382935   .8027781    -9.20   0.000    -8.960192   -5.805678
------------------------------------------------------------------------------
```

Hence, N_c = 500, SSE_c = 7382.85, DFE_c = 497.

We now compute the incremental F:

$$F_{K+1, N_1+N_2-2K-2} = \frac{(SSE_c - SSE_u) * (N_1 + N_2 - 2K - 2)}{SSE_u * (K+1)} = \frac{(7383 - 6770) * 494}{6770 * 3} = 14.91$$

The critical value for an F with d.f. = 3, 494 is only about 2.10. Therefore, we reject the null hypothesis: coefficients are not the same for both blacks and whites. Just from “eyeballing” the coefficients, it appears that both education and years of job experience have smaller effects on blacks than on whites.

Conclusion. Both compositional differences and differences in variable effects contribute to income differences between blacks and white. The descriptive statistics reveal that blacks have lower levels of education and job experience. These lower levels of education and experience, combined with the apparently smaller effects of the education and experience they do have, make black income lower than white income.

Remember our earlier caution, however. We do not know which effects significantly differ across populations. Indeed, it may not even be an effect of an IV that is different; it could be that the intercepts significantly differ. Subsequent handouts will discuss alternative and more flexible ways for making comparisons across groups.

More generally, the researcher should be aware that both differences in composition and differences in independent variable effects could be important when trying to explain why differences in outcomes exist across groups.

Appendix. Corresponding SPSS Commands.

In SPSS, to get the subgroup models you could give commands like

```
REGRESSION
  /SELECT= black EQ 0
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT income
  /METHOD=ENTER educ jobexp  .

REGRESSION
  /SELECT= black EQ 1
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT income
  /METHOD=ENTER educ jobexp  .
```

Or, you could do something like

```
Sort cases by black.
Split file by black.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT income
  /METHOD=ENTER educ jobexp  .
Split file off.
```

With the latter command, the key part of the output would be

ANOVA[b]

BLACK	Model		Sum of Squares	df	Mean Square	F	Sig.
0 white	1	Regression	18361.989	2	9180.995	620.066	.000[a]
		Residual	5878.170	397	14.806		
		Total	24240.159	399			
1 black	1	Regression	4924.273	2	2462.136	267.798	.000[a]
		Residual	891.817	97	9.194		
		Total	5816.090	99			

a. Predictors: (Constant), JOBEXP, EDUC

b. Dependent Variable: INCOME