Version Differences in the LFS

DataFirst

December 6, 2011

February 2000 - v1.1

The South African February 2000 LFS dataset was originally released in 2001 as 3 data files (household, worker and general). A second version was downloaded from the Statistics South Africa website on 11 August 2011 by DataFirst. This version differed slightly from the originally obtained release in the following ways:

1. The suffix "_Feb2000" no longer appears on all variable names
2. Year and Month variables were added
3. Variable labels were altered. Previously, all variable labels were literal questions. Now the variable labels describe the variables.

September 2000 - v2.1

The South African September 2000 LFS dataset was originally released in 2001 as 4 data files (household, worker, person and stratum_psu). A second version was downloaded from the Statistics South Africa website subsequent to that in March 2006 by DataFirst. This version differed slightly from the originally obtained release. Most notably, weights were recast to reflect population estimates released in February 2005. This version was also benchmarked to the 2001 South African census (whereas previously it had been benchmarked to the 1996 South African census). As a result, the weight variables in each data file differ between versions 1.0 and 2.0. The second version (version 2.0) also has several extra observations. The source of these extra data is unclear. Specifically,

- 31 extra observations are in the household data file
- 1 extra observation are in the person data file
- 18 extra observations are in the worker data file

A third version (version 2.1) was downloaded by DataFirst on 11 August 2011 as 3 data files (the other three data files subsumed the, originally separately released, stratum_psu datafile) which differed slightly from version 2.0 in the following ways:

1. The suffix "_Sep2000" no longer appears on all variable names
2. Year and Month variables were added
3. Variable labels were altered. Previously, all variable labels were literal questions. Now the variable labels describe the variables.
4. A number of variables were renamed (beyond dropping the suffix _Sep2000). For example, Q710Lght_Sep2000 in version 2.0 is now Q710Ligh in version 2.1. This could prove confusing when comparing between current and previous versions of the data file(s).

Version 2.1 and 2.0 also have some substantive differences:

Household/general data file

The most significant of these is the apparent switch of one of the variable names and labels in the household/general data files. To clarify, Question 7.25 in the LFS Household questionnaire pertains to proximity to transport. Version 1.1 has entries for question 7.25a that relate respondent proximity to trains, but in the later version this variable relates proximity to taxis. The same is true for question 7.25b (i.e. the converse is also true). Summarily, the data relating respondent proximity to trains and taxis has been muddled up in version 2.0. This mistake would seem to be within version 2.0 only, as version 2.1 agrees with the original (version 1.0).

Furthermore, the variable reflecting the way in which the household receives mail (in the household data file) in version 2.0 has no value labels and does not match up with the values in version 2.1. Version 2.1 aggregates the value labels of version 2.0 in groups of 10. For example, in version 2.0 there are variables labelled 11, 12 and 19, which are grouped into decades in version 2.1. So entries of 11, 12 and 19 all take on the value 1 in the later version. In version 2.0, the sum of the number of observations within decades equals the sum of observations equal to 1 in version 2.0 (value label: "Delivered to the dwelling"). It is unclear as to the source of the distinction between these variables (it may be arbitrary, an artifact of ASCII to STATA conversion for example).

February 2001 - v2.0

The South African February 2001 LFS dataset was originally released in September 2001 as 4 data files (household, worker and general and stratum_psu). A second version was downloaded from the Statistics South Africa website as 3 data files (incorporating the previously seperate stratum_psu data file) on 11 August 2011. This version differed slightly from the originally obtained release in the following ways

1. No "_Feb2001" suffix attached to variable names

2. Year and Month variables included

3. Variable labels were altered. Previously, all variable labels were literal questions. Now the variable labels describe the variables.

4. A number of the variables have been renamed slightly. For example "C_Gender_Feb2001", in the worker data file, version 1.0, is now simply "Gender" in version 2.0

There are also a few substantive differences between the old and new versions of the dataset:

Worker data file

Sector: The variable representing employment sector has had a number of observations that were previously "Unspecified" recoded into various other categories. There are 91 differences between versions in total.

Industry: The variable representing "Main industry" has been recoded for a number of observations. Most importantly, the variable value label previously denoted by "Other" is now labelled "12". Some of the observations have different values to previous versions. There are 316 differences between versions in total.

Main Occupation: The variable representing "Main occupation" has been recoded for a number of observations. Most importantly, the variable value label previously denoted by "Occupation not adequately defined" is now labelled "91". Some of the observations have different values than before. There are 285 differences between versions in total.

Employment Status: The variables representing employment status have several substantive differences. These two variables reflect two definitions, "narrow" and "expanded", of employment status. Some observations that were previously defined as having one particular status (within both variables!) are now defined as another. Official employment status (the "narrow" definition, STATUS1) has 359 differences, whereas expanded employment status (STATUS2) has 583 differences.

Household/general data file

The household/general file in version 2.0 has 13 extra observations, 3 each from the Free State, Northwest and Gauteng provinces and 4 from Limpopo. Each of these observations are coded missing or unspecified for each variable except Province and Household Weight.

September 2001 - v2.1

The South African September 2001 LFS dataset was originally released in March 2002 as 4 data files (household, worker, person and stratum_psu). A second version was downloaded from the Statistics South Africa website subsequent to that in January 2006, which differed slightly from the originally obtained release in that it had weights that were aimed at reflecting the February 2005 population estimates. This version was benchmarked to the 2001 South African census (whereas previously, it had been benchmarked to the 1996 South African census). As a result, the weight variables in each data file differ between versions 1.0 and 2.0.

A third version (version 2.1) was downloaded by DataFirst on 11 August 2011, which differed slightly from version 2.0 in the following ways:

1. No suffix "_Sep2001" attached to the variable names
2. Year and Month variables included
3. Variable labels were altered. Previously, all variable labels were literal questions. Now the variable labels describe the variables.
4. A number of variables were renamed. This could prove confusing when comparing between current and previous versions of the data file.

Version 2.1 and 2.0 also have some substantive differences:

Worker data file

Sector: The variable representing employment sector has had a number of observations recoded into various other categories. There are 40 differences between versions.

Industry: The variables representing "Main industry" (one is specific, the other more general) have been recoded for a number of observations. There are 151 differences between versions for the general (14 unique values) industry variable and 40 differences between versions for the specific (191 unique values) industry label. This would suggest that the differences between the more general industry variables are not driven exclusively by changes to observations in the specific industry variable entry.

Main Occupation: The variables representing "Main occupation" (one is specific, the other more general) have been recoded for a number of observations. There are 129 differences between versions for the general (13 unique values) occupation variable and 40 differences between versions for the specific (371 unique values) occupation variable. This would suggest that the differences between general occupation variable are not driven exclusively by changes to observations in the specific occupation variable entry.

Employment Status: The variables representing employment status have several substantive differences. The two variables reflect two definitions, narrow and expanded, of employment status. Some observations that were previously defined as having one particular (within both variables!) are now defined as another. Official employment status (STATUS1) has 253 differences, whereas expanded employment status has 524 differences.

Household/general data file

The household/general file in version 2.0 has 17 extra observations. Each of these observations are coded missing or unspecified for each variable except Province and Household Weight.

February 2002 - v2.0

The South African February 2002 LFS dataset was originally released in September 2002 as 4 data files (death/mortality, worker, general and stratum_psu). A second version was downloaded from the Statistics South Africa website as 3 data files (incorporating the previously separate stratum_psu data file) on 11 August 2011. This version differed slightly from the originally obtained release in that it had weights that were aimed at reflecting population estimates released in February 2005. This version was also benchmarked to the 2001 South African census (whereas previously, it had been benchmarked to the 1996 South African census). As a result, the weight variables differ slightly between data files in versions 1.0 and 2.0. Furthermore, the second version (version 2.0) had several extra observations, the source of which is unclear in the documentation provided by Statistics South Africa. There are several other differences between this version of the dataset and the previous one

1. No suffix "_Feb2002" attached to the variable names
2. Year and Month variables included
3. Variable labels were altered. Previously, all variable labels were literal questions. Now the variable labels describe the variables.
4. A number of the variables have been renamed slightly. For example "DEATHNR", in the mortality data file, version 1.0, is now "DEATNR" in version 2.0

There are also a few substantive differences between the old and new datasets:

Mortality data file

"Q66AGEDI": This variable, which reflects the age of the deceased, has been changed into what appears to be an uncoded categorical. Previously, this variable reflected the exact age of the deceased. Now, it instead reflects an age range.

Person and worker data files

The person and worker data files in version 2.0 have 2 extra observations, both from the same household in Kwazulu-Natal.

September 2002 - v2.1

The South African September 2002 LFS dataset was originally released in March 2003 as 5 data files (household, worker, migrant, person and stratum_psu). A second version was downloaded from the Statistics South Africa website subsequent to that in July 2006. This version differed slightly from the originally obtained release. Most notably, weights were recast to reflect population estimates released in February 2005. This version was also benchmarked to the 2001 South African census (whereas previously it had been benchmarked to the 1996 South African census). As a result, the weight variable(s) in versions 1.0 and 2.0 are different.

A third version (version 2.1) was downloaded by DataFirst on 11 August 2011, which differed slightly from version 2.0 in the following ways:

1. No suffix "_Sep2002" attached to the variable names
2. Year and Month variables included
3. Variable labels were altered. Previously, all variable labels were literal questions. Now the variable labels describe the variables.
4. A number of the variables have been renamed slightly This could prove confusing when comparing between current and previous versions of the data file.

Version 2.1 and 2.0 also have some substantive differences:

Worker data file

Sector: The variable representing employment sector has had a number of observations recoded into various other categories. There are 33 differences between versions.

Industry: The variables representing "Main industry" (one is specific, the other more general) have been recoded for a number of observations. There are 175 differences between versions for the general (14 unique values) industry variable and 23 differences between versions for the specific (191 unique values) industry label. This would suggest that the differences between the more general industry variables are not driven exclusively by changes to observations in the specific industry variable entry.

Main occupation: The variables representing "Main occupation" (one is specific, the other more general) have been recoded for a number of observations. There are 147 differences between versions for the general (13 unique values) occupation variable and 23 differences between versions for the specific (377 unique values) occupation variable. This would suggest that the differences between general occupation variable are not driven exclusively by changes to observations in the specific occupation variable entry.

Employment Status: The variables representing employment status have several substantive differences. The two variables reflect two definitions, narrow and expanded, of employment status. Some observations that were previously defined as having one particular (within both variables!) are now defined as another. Official employment status (STATUS1) has 229 differences, whereas expanded employment status (STATUS2) has 353 differences.

Household/general data file:

The household/general file in version 2.0 has 55 extra observations, which appear to come from all around the country, but mostly from Gauteng (37 of the "extra" observations are from Gauteng)

March 2003 - v2.0

The South African March 2003 LFS dataset was originally released in September 2003 as 3 data files (worker, person and stratum_psu). A second version was downloaded from the Statistics South Africa website in August 2011 by DataFirst. This version differed slightly from the originally obtained release. Most notably, weights were recast to reflect population estimates released in February 2005. This version was also benchmarked to the 2001 South African census (whereas previously it had been benchmarked to the 1996 South African census). As a result, the weight variables in each data file differ between versions 1.0 and 2.0. The stratum_psu data file has also been subsumed into the other data files.

The datasets also differ in the following ways,

1. No suffix "_Mar2003" in some variable names
2. Variable labels were altered. Previously, all variable labels were literal questions. Now the variable labels describe the variables.
3. A number of variables were renamed (beyond dropping the suffix "_Mar2003"). This could prove confusing when comparing between current and previous versions of the data file.

Version 1.0 and 2.0 also have some substantive differences:

Worker data file

Sector: The variable representing employment sector has had a number of observations recoded into various other categories. There are 24 differences between versions.

Industry The variables representing "Main industry" (one is specific, the other more general) have been recoded for a number of observations. There are 135 differences between versions for the general (14 unique values) industry variable and 20 differences between versions for the specific (199 unique values) industry label. This would suggest that the differences between the more general industry variables are not driven exclusively by changes to observations in the specific industry variable entry.

Main Occupation: The variables representing "Main occupation" (one is specific, the other more general) have been recoded for a number of observations. There are 131 differences between versions for the general (13 unique values) occupation variable and 20 differences between versions for the specific (370 unique values) occupation variable. This would suggest that the differences between general occupation variable are not driven exclusively by changes to observations in the specific occupation variable entry.

Employment Status: The variables representing employment status have several substantive differences. The two variables reflect two definitions, narrow and expanded, of employment status. Some observations that were previously defined as having one particular (within both variables!) are now defined as another. Official employment status (STATUS1) has 172 differences, whereas expanded employment status (STATUS2) has 260 differences.

September 2003 - v2.1

The South African September 2003 LFS dataset was originally released in March 2004 as 4 data files (house, migrant, worker and person). A second version was downloaded from the Statistics South Africa website in July 2006 by Datafirst. This version differed slightly from the originally obtained release in a number of ways. Most notably, weights were recast to reflect population estimates released in February 2005. This version was also benchmarked to the 2001 South African census (whereas previously it had been benchmarked to the 1996 South African census). As a result, the weight variable(s) in each data file differ between versions 1.0 and 2.0.

A third version (version 2.1) was downloaded by DataFirst on 11 August 2011, which differed slightly from version 2.0 in the following ways,

1. No suffix "_Sep2003" in some variable names

2. Variable labels were altered. Previously, all variable labels were literal questions. Now the variable labels describe the variables.

3. A number of variables were renamed (beyond dropping the suffix "_Sep2003"). This could prove confusing when comparing between current and previous versions of the data file.

Version 2.0 and 2.1 have some substantive differences:

Household/general data file

The household/general file in version 2.1 has 10 extra observations, all of which are in Gauteng

Worker data file

The worker file in version 2.1 has 19 extra observations, all of which come from the abovementioned dwellings in Gauteng

March 2004 - v2.0

The South African March 2004 LFS dataset was originally released in September 2004 as 3 data files (household, person and worker). A second version was downloaded from the Statistics South Africa website in August 2011 by DataFirst. This version differed slightly from the originally obtained release. Most notably, weights were recast to reflect population estimates released in February 2005. This version was also benchmarked to the 2001 South African census (whereas previously, it had been benchmarked to the 1996 South African census). As a result, the weight variables in each data file differ between versions 1.0 and 2.0.

September 2004 - v2.0

The South African September 2004 LFS dataset was originally released in March 2005 as 4 data files (worker, person, migrant and household). A second version was downloaded from the Statistics South Africa website in August 2011 by DataFirst. The only differences between these two versions is that 2.0 includes year and month variables and version 1.0 represented all variable names in lower case.

March 2005 - v1.1

The South African March 2005 LFS dataset was originally released in September 2005 as 2 data files (person and worker). A second version was downloaded from the Statistics South Africa website subsequent to that in August 2011. This version differs from the original release in that all data files now contain year and month variables.

September 2005 - v1.1

The South African September 2005 LFS dataset was originally released in March 2006 as 2 data files (person and worker). A second version was downloaded from the Statistics South Africa website subsequent to that in August 2011. This version differs from the original release in that all data files now contain year and month variables.

March 2006 - v1.1

The South African March 2006 LFS dataset was originally released in September 2006 as 2 data files (person and worker). A second version was downloaded from the Statistics South Africa website subsequent to that in August 2011. This version differs from the original release in that all data files now contain year and month variables.

September 2006 - v1.1

The South African September 2006 LFS dataset was originally released in March 2007 as 2 data files (person and worker). A second version was downloaded from the Statistics South Africa website subsequent to that in August 2011. This version differs from the original release in that all data files now contain year and month variables.

March 2007 - v1.1

The South African March 2007 LFS dataset was originally released in September 2007 as 2 data files (person and worker). This version appears unchanged as of August 2011.

September 2007 - v2.1

The South African September 2007 LFS dataset was originally released in March 2007 as 2 data files (person and worker), before being immediately retracted due to a number of errors identified by Statistics South Africa. A second version was made available shortly thereafter by Statistics South Africa (version 2.0). A third version was then downloaded by DataFirst in November 2009. This version (version 2.1), the worker data file in particular, differs from version 2.0.

Specifically, the variable "Q31inHHP" (which is a categorical reflecting whether or not the respondent is supported by members of his or her household) differs substantively between versions. In version 2.0, the data shows roughly 13,388 observations responding "No" (a value of 2) and roughly 26,287 assigned the

value label "N/A" (data value of 8). This reverses in version 2.1, with 26,287 respondents now listed as "No" and 13,388 listed as "3" for that variable. The value labels of "N/A" and "Unspecified" are no longer in the data file and the actual data values have also changed to 3 and 4. It is assumed that the value 9, previously labelled as "Unspecified", is now simply equivalent to the unlabelled data value "4" (which is no longer labelled at all). This is assumed because the count of observations of that value is the same between versions of the data file, although this may be false given that the observation count reverses for two other values in the same variable. The original metadata specifies that data values of 8 and 9 should be labelled "N/A" and "Unspecified" respectively. The change in the data between versions, then, looks likely to be an error of some kind.

A third version was downloaded by DataFirst in August 2011, which is the same as version 2.1.