AN INTRODUCTION TO THE UT/DLS: MICRODATA ANALYSIS SUBSETTING (SDA@UOFT)

A RESOURCE TO HELP YOU LEARN AND USE THE UT/DLS

PREPARED BY:

LYNDA GAGNE,
LINDSAY TEDDS[1],
JENNIFER SULLIVAN,
AND
KATE BERNIAZ
SCHOOL OF PUBLIC ADMINISTRATION
UNIVERSITY OF VICTORIA

[1] Contact author: Email: ltedds@uvic.ca, Phone: 250-721-8068

Table of Contents

Introduction

This user guide introduces you to the UT/DLS: Microdata Analysis and Subsetting service that is licensed by the University of Victoria Libraries from the University of Toronto Data Library Service (UT/DLS). This service is based on the Survey Documentation & Analysis (SDA) software developed by the University of California, Berkeley. SDA is a series of software programs that:

- presents documentation associated with survey data sets;
- facilitates Web-based analysis of survey data, and
- includes procedures for creating customized subsets of data that can be downloaded for use in Excel and other statistical software packages.

The UT/DLS service provides access to hundreds of survey and census data sets and documentation, including those from the Statistics Canada Data Liberation Initiative (DLI) collection.[2] A key advantage to the UT/DLS service is that it allows for web-based analysis of a variety of Statistics Canada and other survey and census data. It can be used free of charge through participating universities library system which means that we do not have to: (1) purchase a costly stand alone statistical program (which often run in the hundreds of dollars); (2) learn a complex stand alone statistical program (e.g. SPSS); and (3) download and transport the data on flash drives or CDs because under the UT/DLS the data is always available on the web.

This introduction covers the basic steps involved in accessing survey data and documentation through the UT/DLS: Microdata Analysis and Subsetting, analyzing the data, and manipulating and downloading the data in various forms.

[2] For overview information on the DLI initiative, please see: http://www.statcan.ca/english/Dli/whatisdli.htm. The Statistics Canada DLI data sets are subject to the DLI License that restricts access to current students, faculty and staff of participating universities for their research and teaching. Use of the data files for commercial purposes is strictly forbidden. Data files may not be used under any circumstance for personal contract activities. For more information about use of Statistics Canada DLI data sets, please read this general licensing agreement: http://www.statcan.ca/english/Dli/caselaw/pdf/dlilicence.pdf. A key feature of this licensing agreement is that users are required to cite Statistics Canada as the source of the date in any published research and to indicate that the results or views expressed are those of the author/authorized user and are not those of Statistics Canada. For more information on how to cite Statistics Canada products, please see: http://www.statcan.ca/english/freepub/12-591-XIE/12-591-XIE2006001.htm.

UT/DLS Basics

We begin by discussing the basics involved in accessing survey documentation and data using the UT/DLS service.

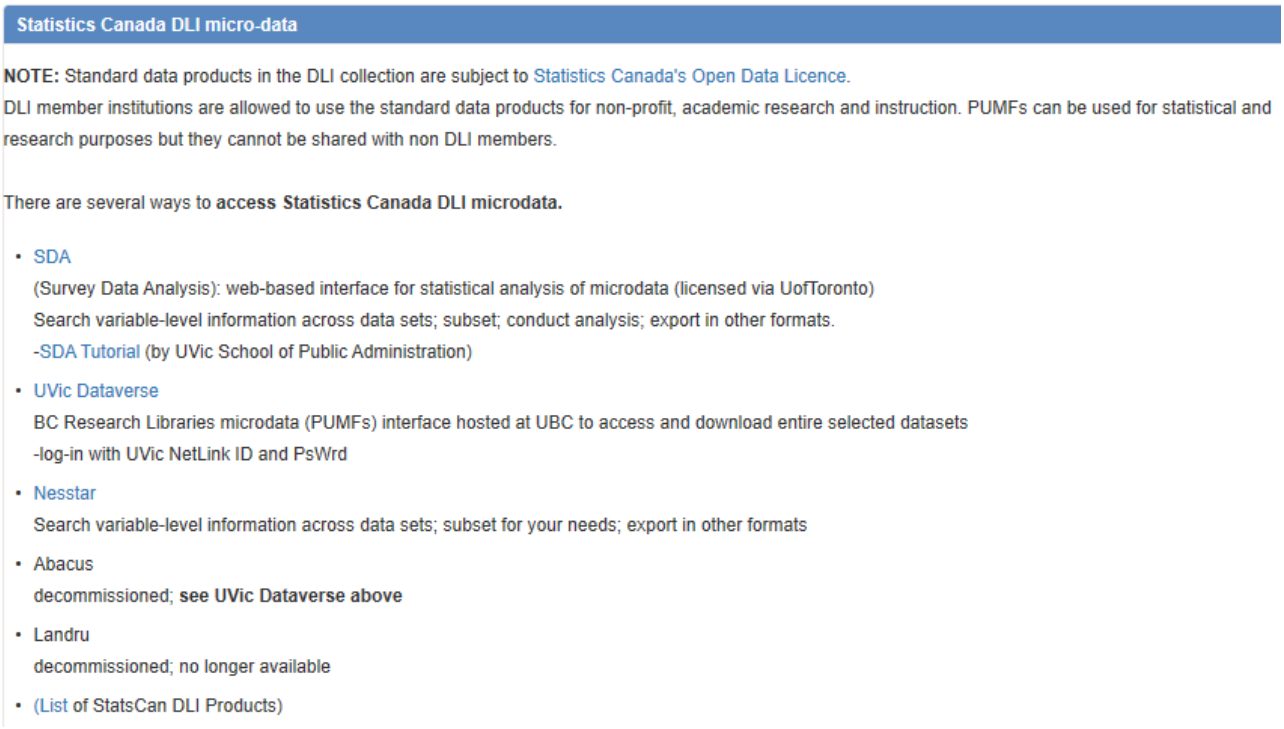
1. Learning Objectives

Upon completion of this section you will be able to:

- Access UT/DLS and its documentation
- Identify available survey data analysis (SDA) programs
- Access data from UT/DLS and examine variable frequencies

2. Accessing UT/DLS: Microdata Analysis and Subsetting Service

To access survey data and documentation licensed by the University of Victoria Libraries go to https://libguides.uvic.ca/socscihumdata



Statistics Canada DLI micro-data

NOTE: Standard data products in the DLI collection are subject to Statistics Canada's Open Data Licence.
DLI member institutions are allowed to use the standard data products for non-profit, academic research and instruction. PUMFs can be used for statistical and research purposes but they cannot be shared with non DLI members.

There are several ways to **access Statistics Canada DLI microdata.**

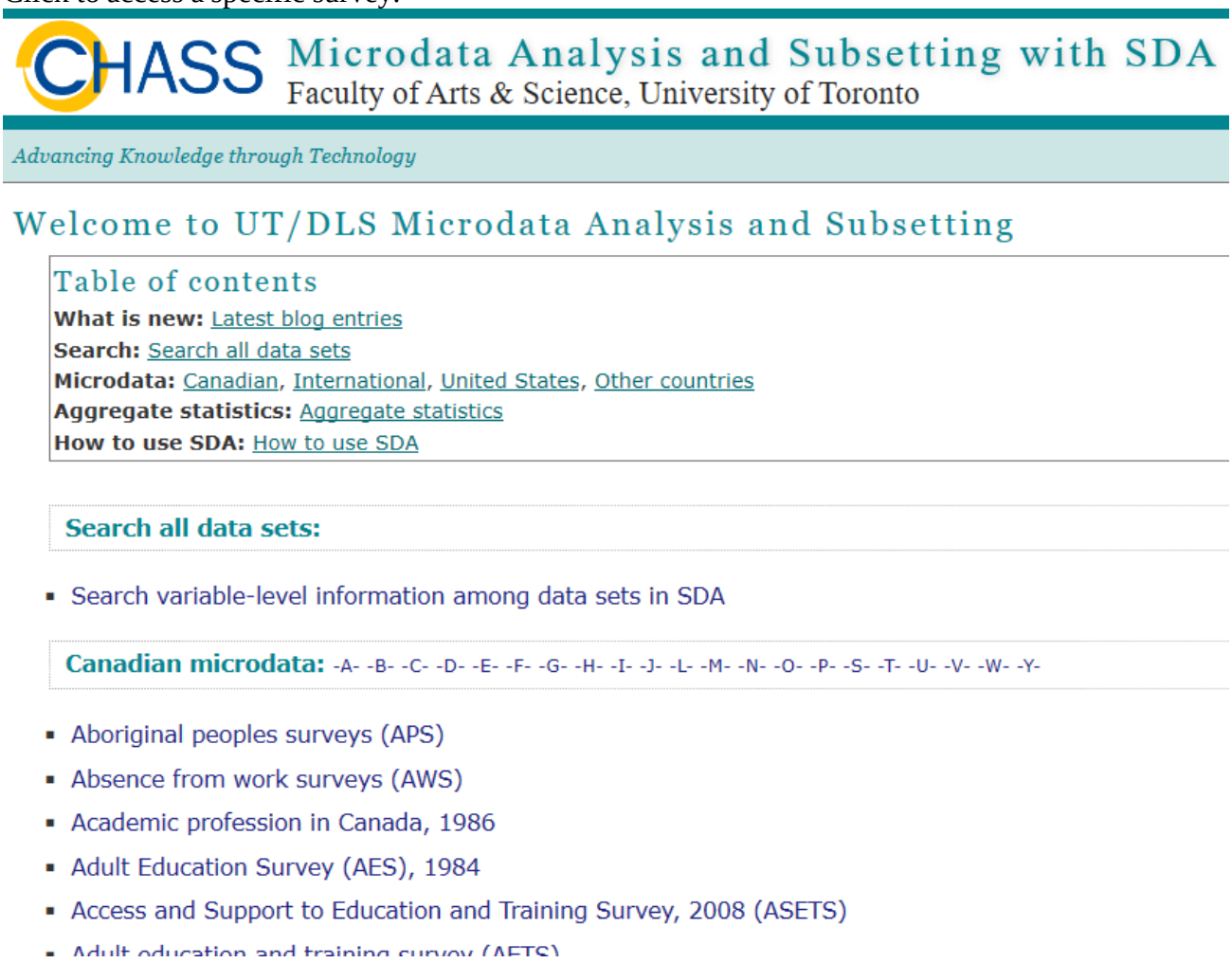
- SDA
  (Survey Data Analysis): web-based interface for statistical analysis of microdata (licensed via UofToronto)
  Search variable-level information across data sets; subset; conduct analysis; export in other formats.
  -SDA Tutorial (by UVic School of Public Administration)
- UVic Dataverse
  BC Research Libraries microdata (PUMFs) interface hosted at UBC to access and download entire selected datasets
  -log-in with UVic NetLink ID and PsWrd
- Nesstar
  Search variable-level information across data sets; subset for your needs; export in other formats
- Abacus
  decommissioned; **see UVic Dataverse above**
- Landru
  decommissioned; no longer available
- (List of StatsCan DLI Products)

[3] Faculty, staff, and students at other participating institutions should consult with their data librarian regarding information on the location of their SDA access page.

You will be taken to a list of surveys available on the server, including, U.S. and International surveys. The surveys available through this survey include but are not limited to:

Canadian Addiction Survey
Canadian Community Health Surveys (CCHS)
Canadian Elections Surveys
Canadian Tobacco Use Monitoring Surveys (CTUMS)
General Social Survey (GSS)
International Adult Literacy and Skills Survey (IALSS)
National Longitudinal Survey of Children and Youth (aka the 'KIDS survey')
Survey of Labour Income Dynamics (SLID)
World Values Survey

Click to access a specific survey:

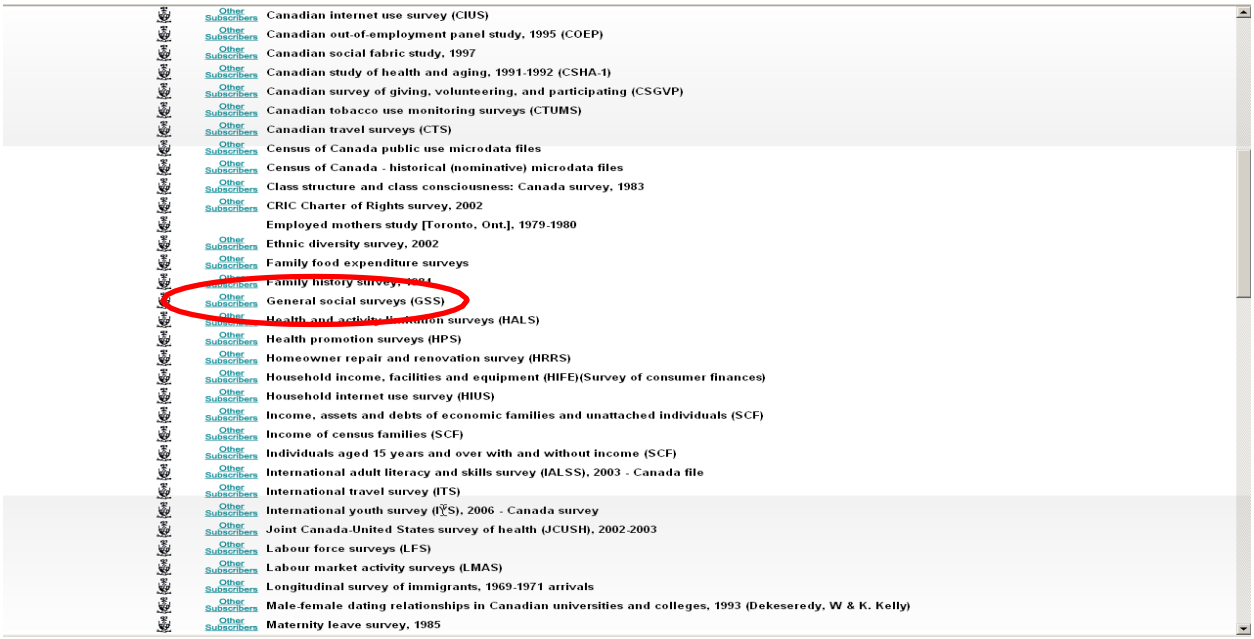


3. Accessing Specific Survey Documentation

A survey's "Documentation" link will take you to a list of all of the available documentation associated with the survey. The available survey documentation will vary between surveys but generally includes: a survey overview, a survey user guide, the survey questionnaire and/or an index of variables.

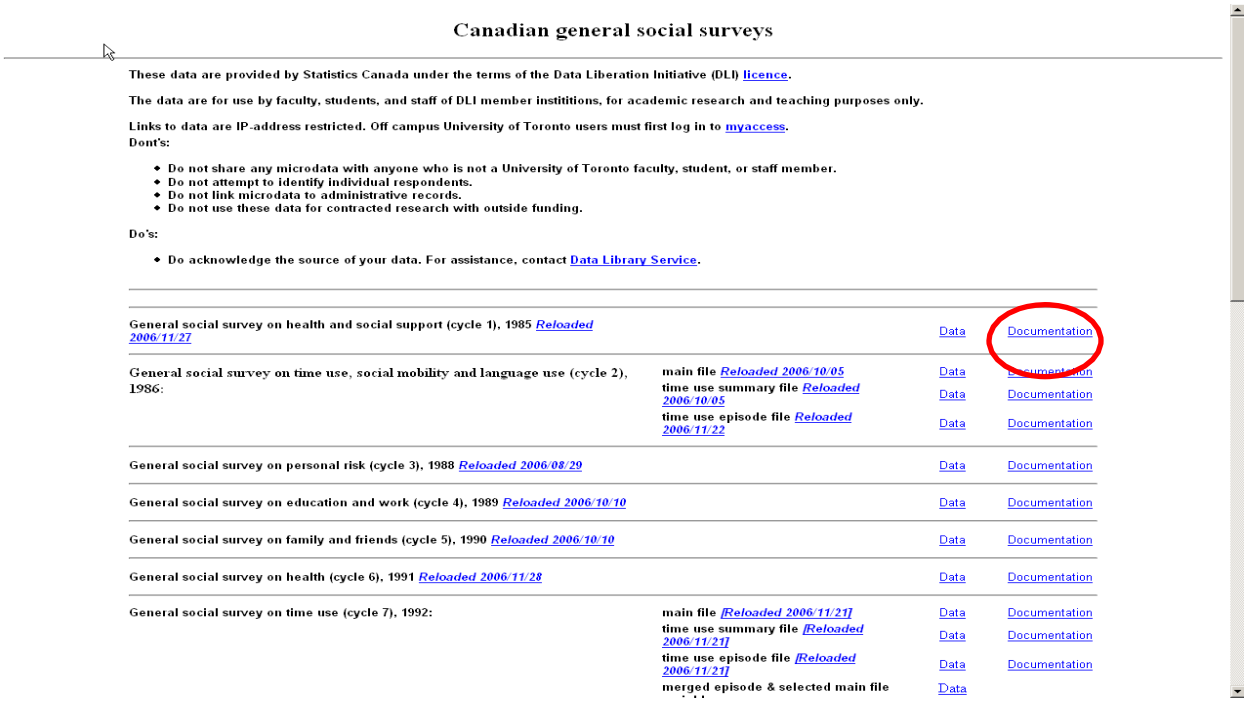
Example: Accessing the General Social Surveys (GSS) Documentation

- ✓ Under "Canadian microdata" left click "Other subscribers" next to "General social surveys (GSS)".

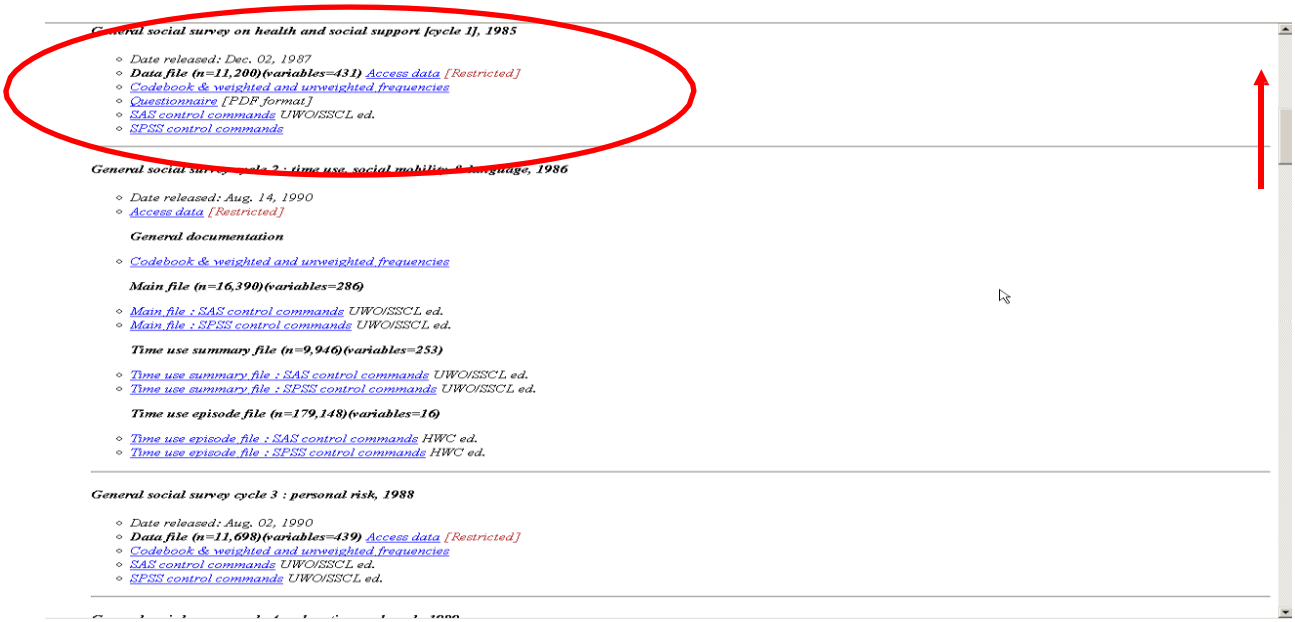


This will take you to a list of all GSS on the server. There are many different cycles of the GSS, each addressing a different topic (e.g. cycle 8 = personal risk; cycle 15 = family history).

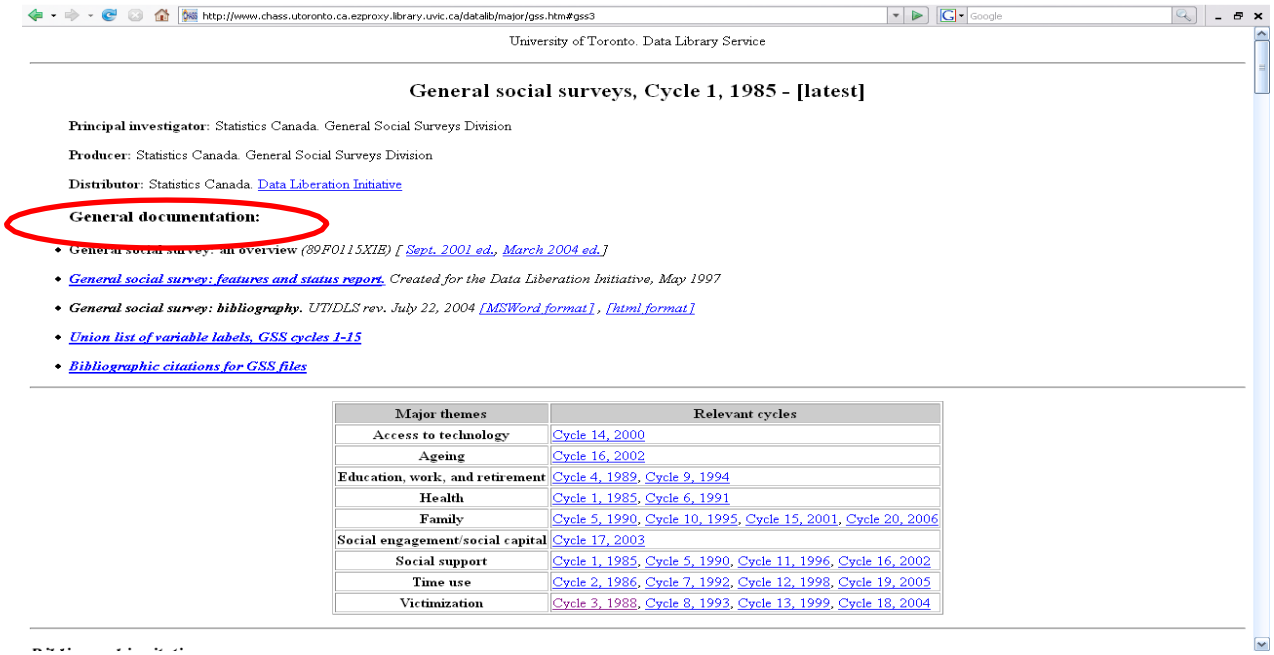
- ✓ Left click "Documentation" next to any survey cycle.



This will take you to the survey documentation page.



Scroll to the top of the page to see the general documentation related to the survey.



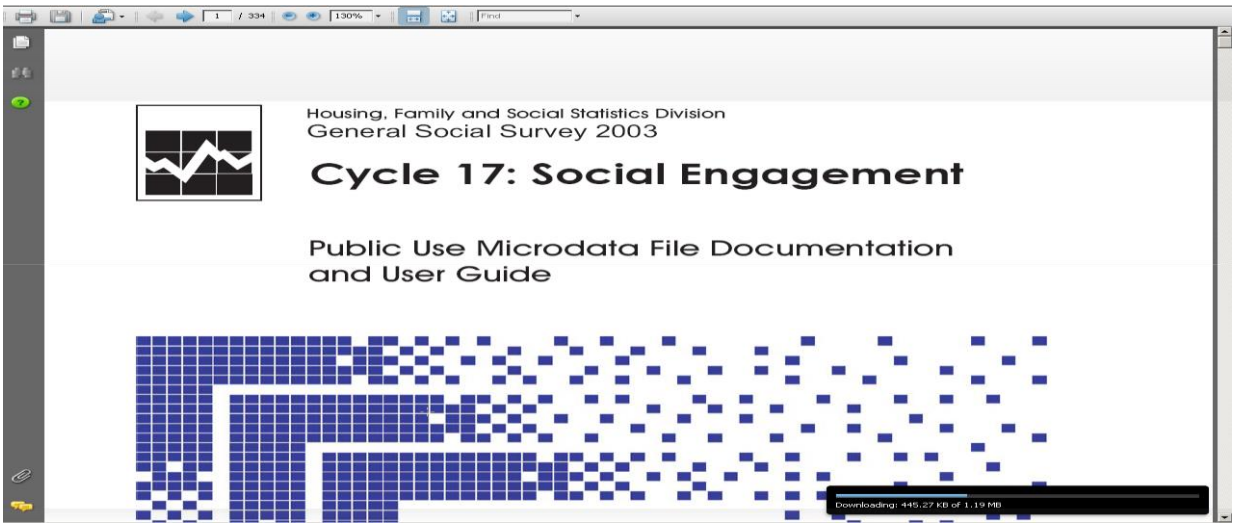
General Documentation available includes:

- A survey overview for the 2001 and 2004 editions of the survey including background, target population, collection methodology and survey content description;
- A Features and Status Report created for the Data Liberation Initiative;
- A General social survey bibliography;
- A list of survey variables for cycles 1-15, and
- Bibliographic Citation format for GSS files.

Browse through these links and familiarize yourself with the information available. Scroll down the page to find additional documentation related to specific survey cycles, including User Guides and Questionnaires.

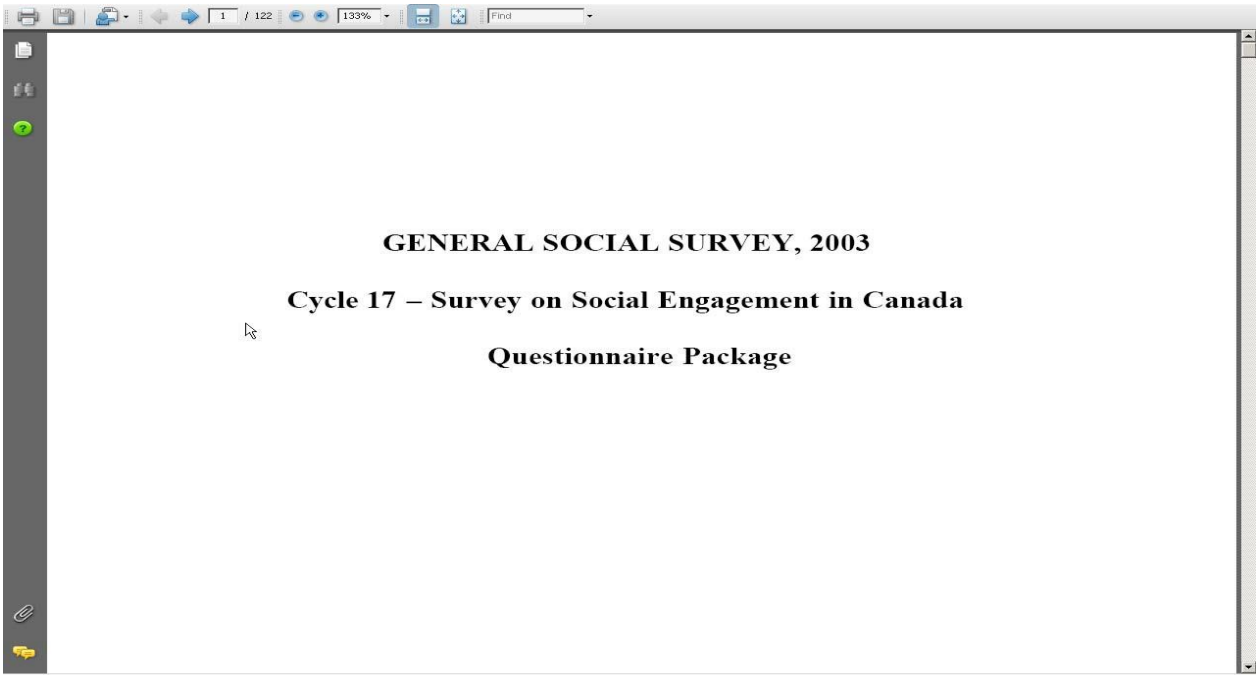
- ✓ Scroll down to "*General social survey cycle 17: social engagement, 2003".*
- ✓ Left click on link next to **"*User Guide*".**

This will take you to the "Public Use Microdata file Documentation and User's Guide" for the 2003 GSS on Social Engagement. The document contains information on the objectives of the GSS, the content and special features of the specific cycle, survey and sample design, and data collection and processing.



- ✓ Go back to the main GSS documentation page (using your browser back button).
- ✓ Under "*General social survey cycle 17: social engagement, 2003",* left click on the link next to "***Questionnaire***"

This will take you to the actual questionnaire used in the GSS cycle 17.



Browse through the documentation available for other surveys that you may be interested.

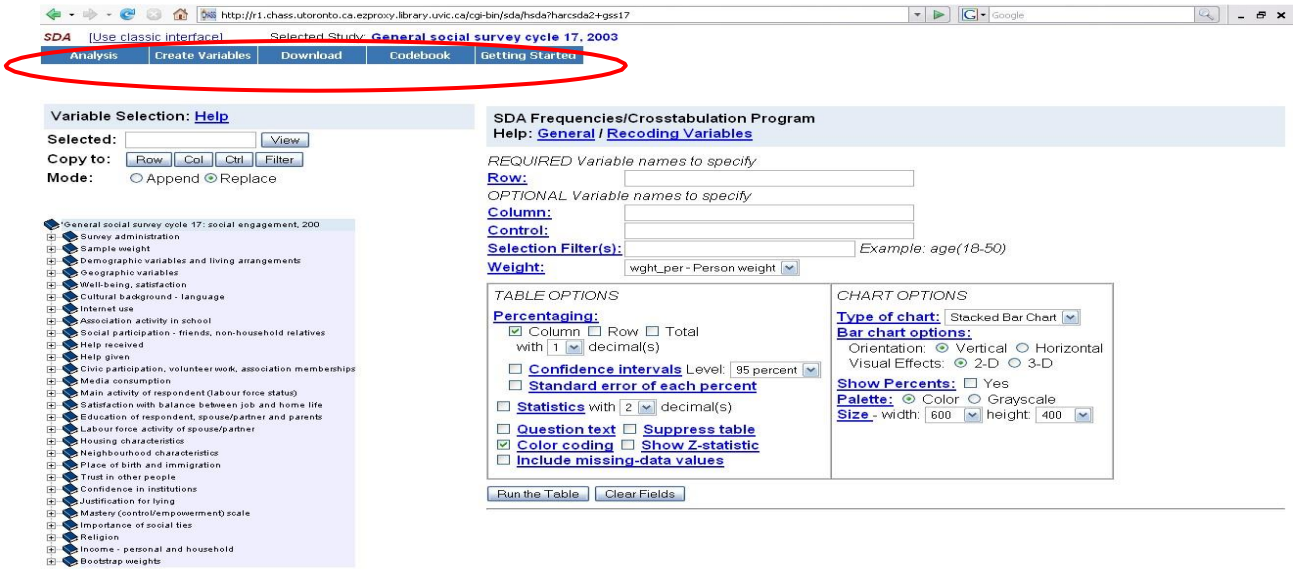
4. Survey Data Analysis Programs

The SDA programs available to analyze survey data are accessed through the "Data" link beside each specific survey. In order to access specific survey data and the analysis programs, from the UT/DLS: Microdata Analysis and Subsetting page:

- ✓ Left click on "Other subscribers" beside the survey you are interested in.

- ✓ Left click the "Data" link.

This link will take you to the main SDA screen. The tool bar at the top of the screen contains the following tabs: Analysis, Create Variables, Download, Codebook and Getting Started.



Scroll through the tabs to view the available programs and functions:

Analysis

- Frequencies or crosstabulation (Default)
- Comparison of means
- Correlation matrix
- Comparison of correlations
- Multiple regression (not covered in this document)
- List values of individual cases
- Logit or probit regression (not covered in this document)

Create Variables

- Compute a new variable
- Recode one or more existing variables, and create a new variable
- List and/or delete the variables created by recoding or computing

Download Files

- Download existing dataset and/or documentation
- Create and download a customized subset of variables

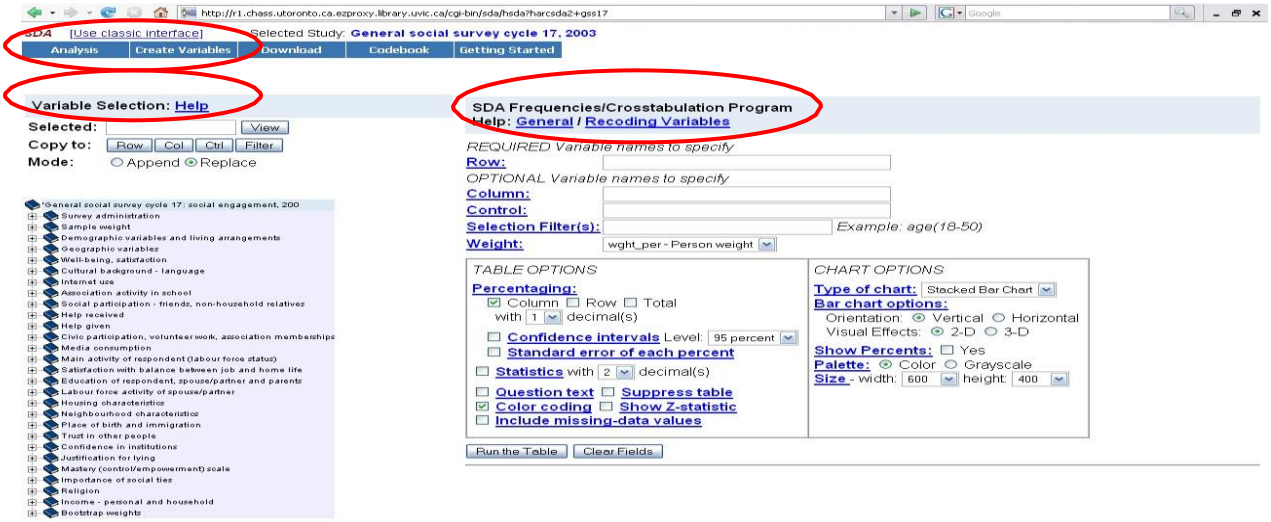
Codebook

- View the full codebook for this dataset (multiple codebooks are sometimes available)

Getting Started

- View help file

On left hand side of the page is displayed the Variable Selection tool which allows you to select variables for the tree menu below and have their names moved over to the program form on the right hand side of the page. The default program form is the SDA Frequencies/Crosstabulation Program. To change the displayed program select the desired program from either the Analysis or Create Variables tabs.

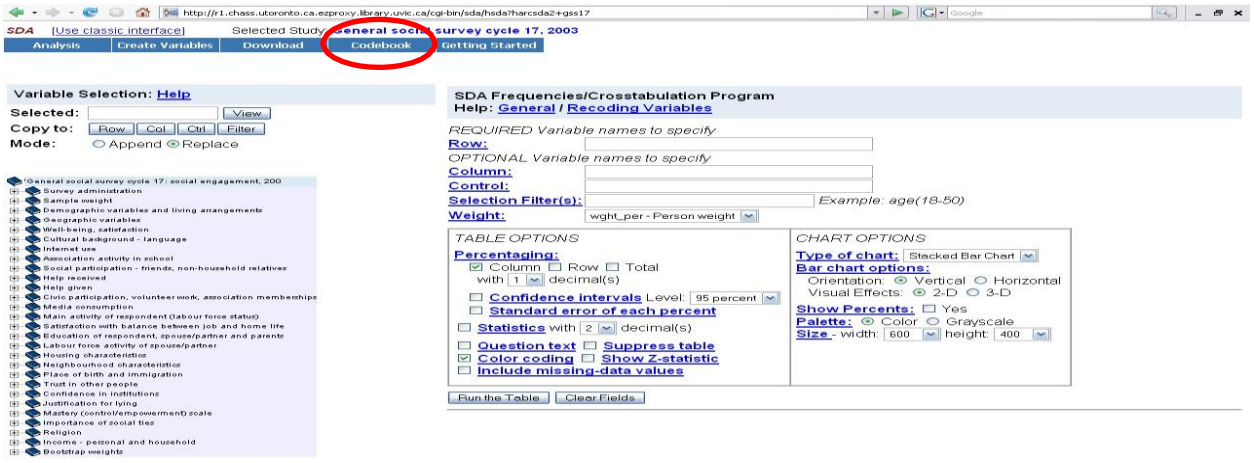


5. Survey Data and Variable Frequencies

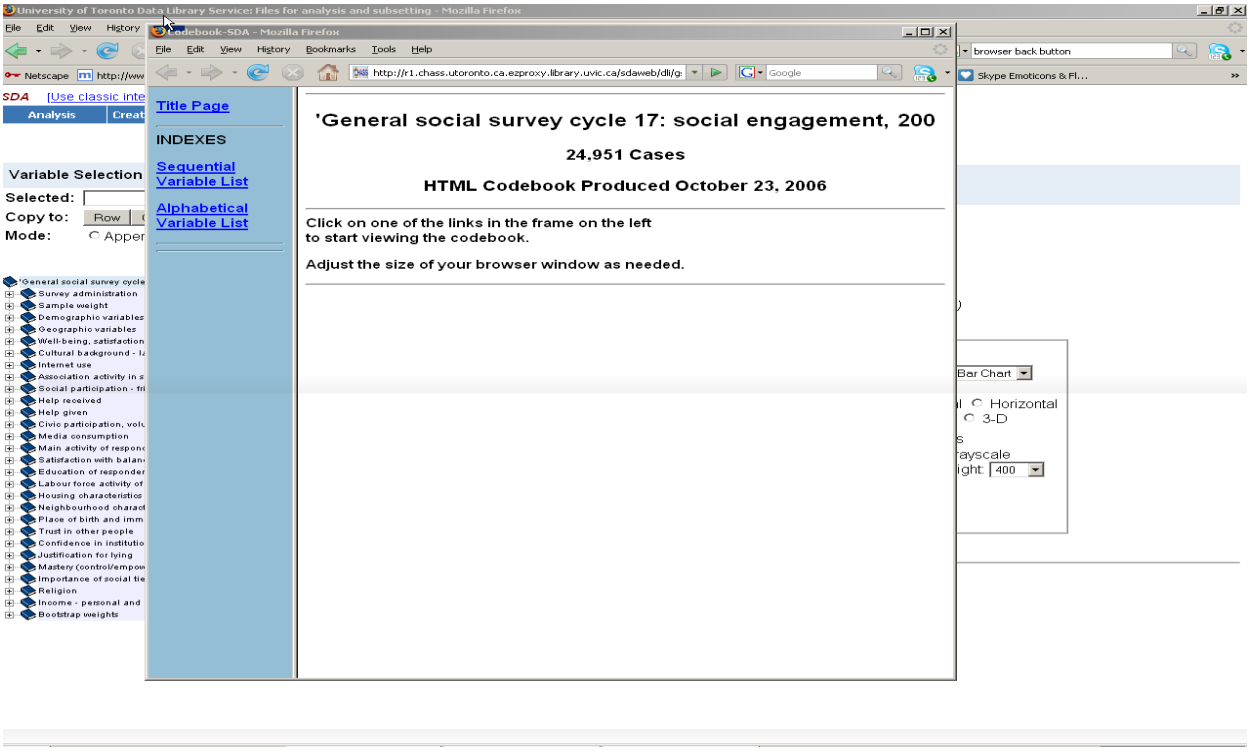
Survey data and specific survey variable frequencies can be viewed using the "Codebook" tab at the top of the main SDA screen

Example: GSS Cycle 17 Variables

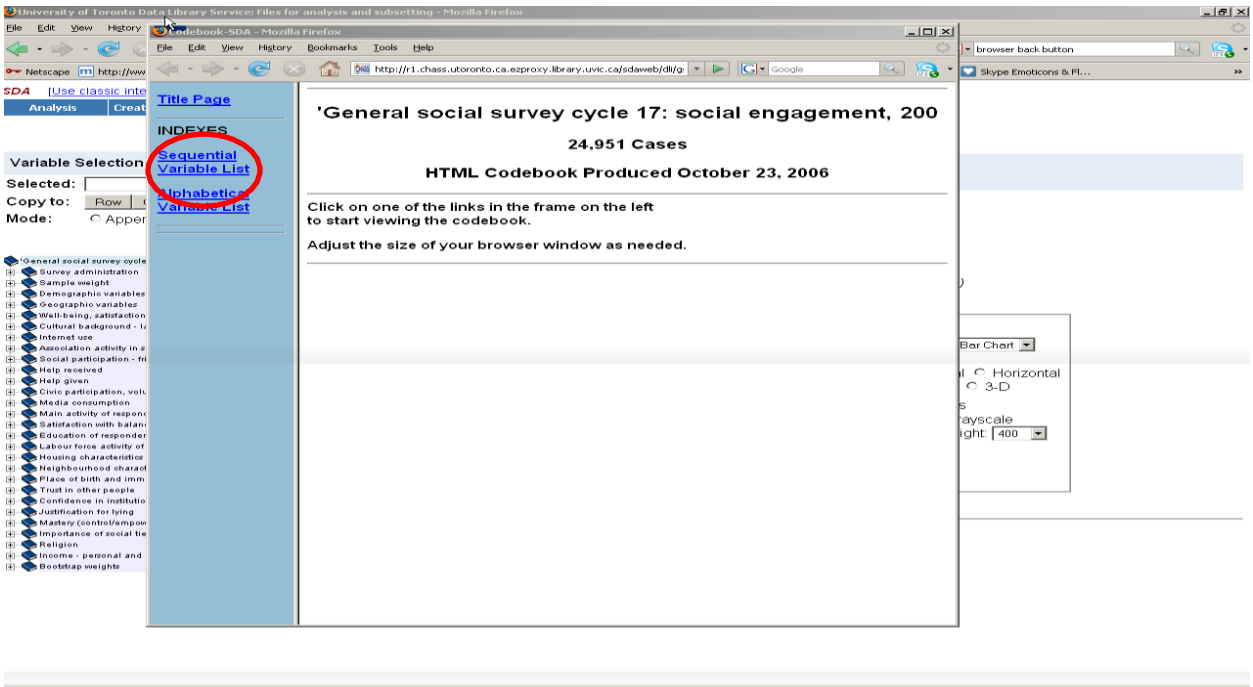
- ✓ Left click on "Other subscribers" beside the GSS.
- ✓ Left click the "Data" link beside General social survey on social engagement (cycle 17), 2003 *Rev. 2006/06/23*
- ✓ Left click on the "Codebook" tab at the top of the page.



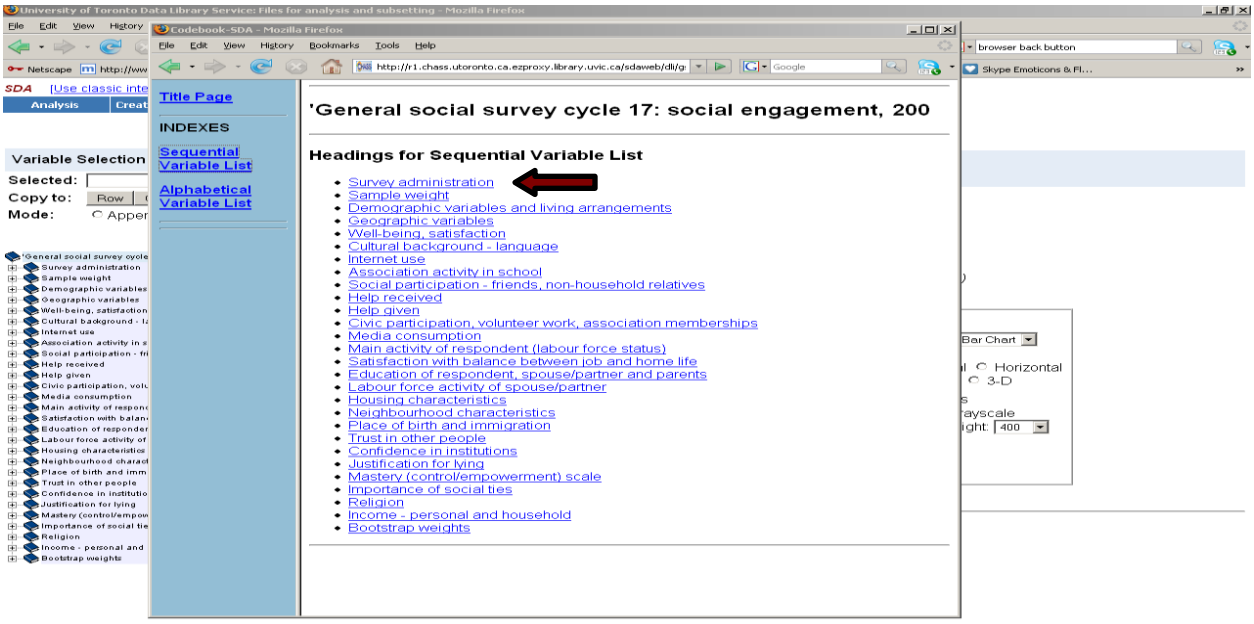
This will open up a new window.



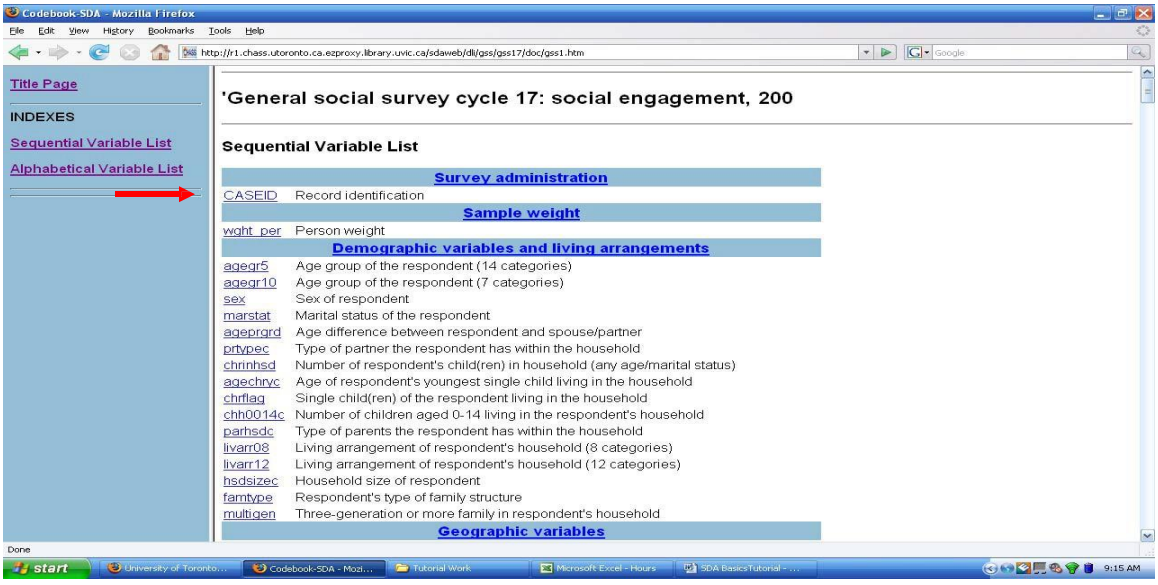
- ✓ Left click on "Sequential Variable List" on the left hand side of the page. This will take you to a list of all of the variable headings used in the survey presented in the order in which they were collected (the "Alphabetical Variable List" presents the survey variables in alphabetical order.



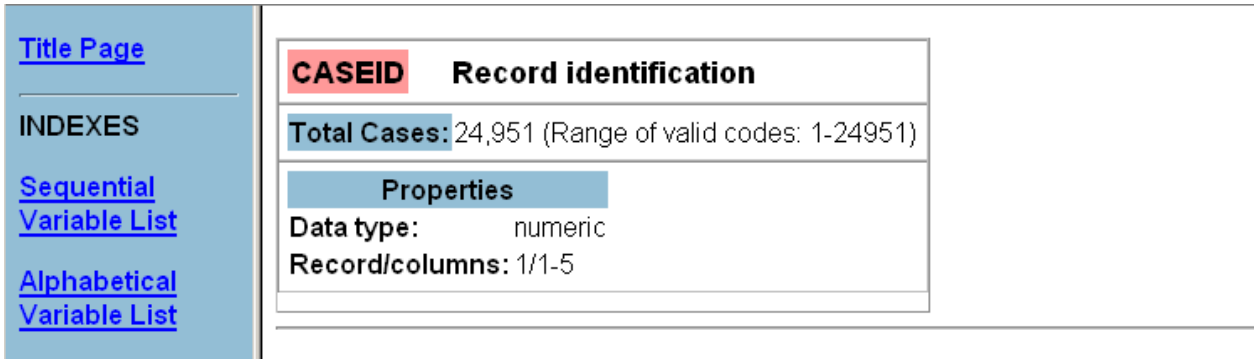
- ✓ Left click on any of the headings. This will take you to an expanded list of all of the variables used in the survey and their associated codes. For example, click on Survey Administration.



This brings you to a list of variables organized by the categories.



- ✓ Left click on any variable on the left hand side of the page. This will take you to a page that summarizes the frequency data for each variable in the survey. For example, if you left click on CASEID, you will see the following result:



✓ The results indicate that there were 24,951 separate records in the survey and that CASEID variable is coded numerically and is up to five digits long.

Use this feature to find out more about survey respondents, for example:

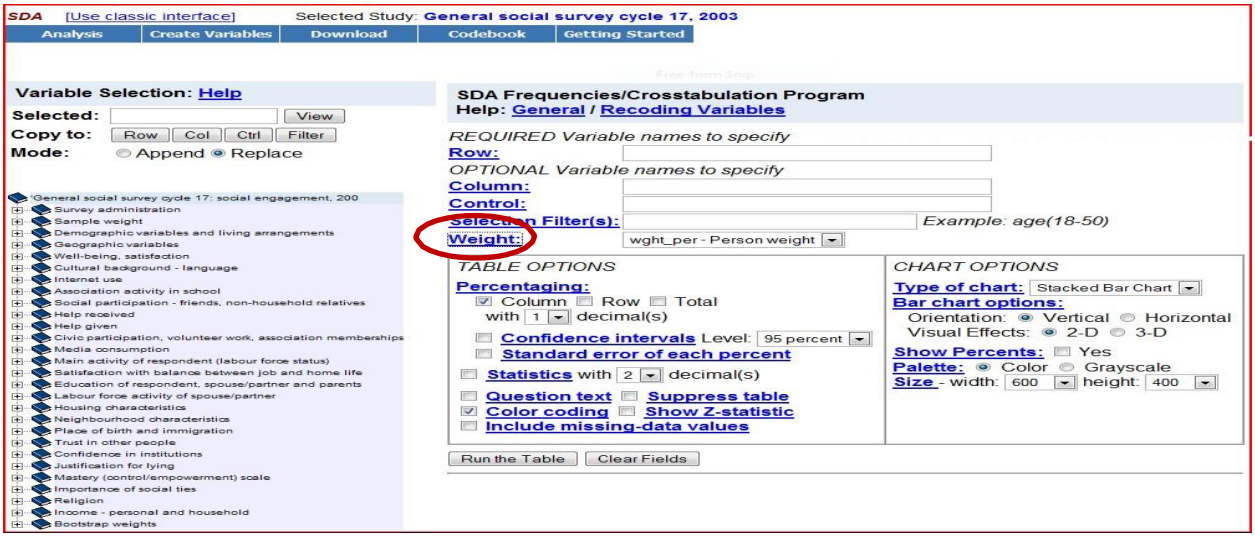
- Age of respondents (variables **agegr5** & **agegr10**)
- Marital status of respondents (variable **marstat**)
- Province of residence of respondents (variable **prv**)
- Respondents self assessed health rating (variable **hal_q110**)
- Respondents main source of stress (variable **mss_q120**)

Please note carefully that the frequency distributions displayed in the Codebook are **unweighted frequencies.**

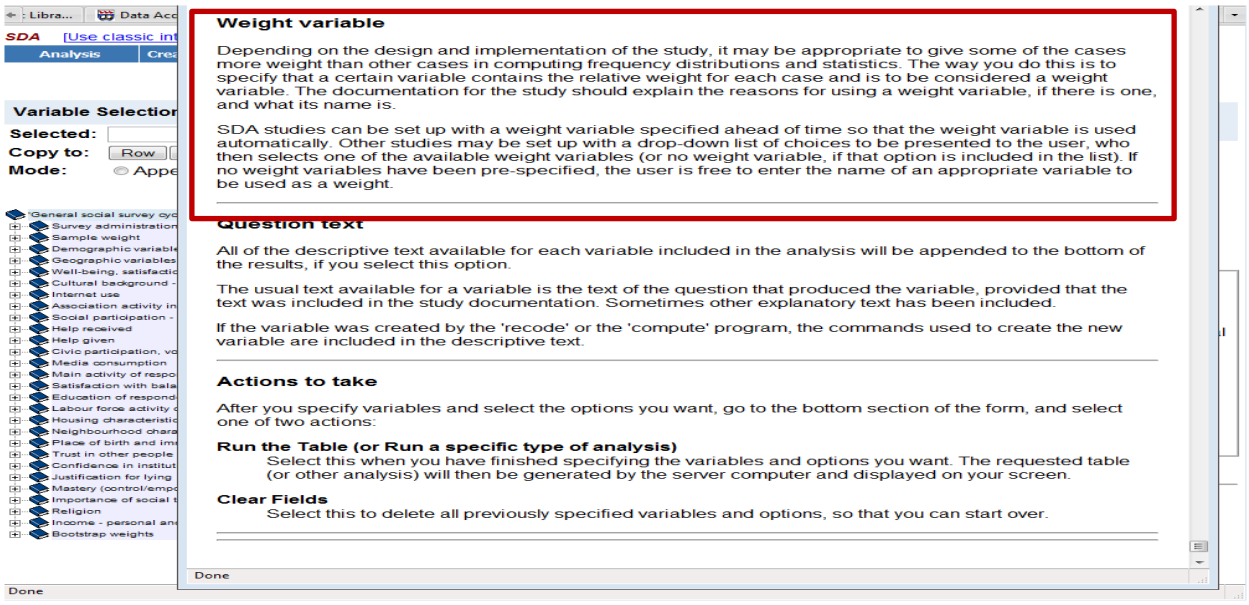
6. Obtaining Help in the UT/DLS tool

If at any time you want more information on the options provided in the UT/DLS tool, you can click on any word in blue to pull up a more detailed description of the feature and/or how to use the feature.

For example, if we wanted to learn more about the using survey weights, we can click on the word "Weight" which is in blue:



and a window pops up that tells us more about this feature.



7. Conclusion

This tutorial covers the material required to access the documentation and data housed on the UT/DLS services and identify the data analysis programs available. In the next section, we will cover how to produce summary statistics and other measures using UT/DLS. To ensure that you understand the basics, please work through the following exercises.

8. Exercises

a) Using the Canadian Community Health Survey cycle 1.1 identify the following information:

i) The number of surveys administered
ii) The self perceived health of respondents
iii) The barriers to improving health and their frequencies

b) b) Using the Survey of Labour and Income Dynamics (SLID) wave 11 identify the following:

i) The percentage of census families who received social assistance (SA) in the reference year
ii) The percentage of census families who received employment insurance (EI) in the reference year
iii) The percentage of economic families who received social assistance (SA) in the reference year
iv) The percentage of economic families who received employment insurance (EI) in the reference year
v) What is the difference between “census family type” and “economic family type”?

9. Answers to Exercises

a) From the UT/DLS: Microdata Analysis and Subsetting page left click on "Other subscribers" next to the "Canadian Community Health Surveys (CCHS)". Left click on "Data" next to "Cycle 1.1, 2000-2001 *[Rev. 08/2003]*". Left click on the "Codebook" tab at the top of the page. This will bring you to the survey codebook. Left click on "Sequential Variable List" on the left had side of the page and then select any variable heading and then any individual variable. This will bring you to the list of all the survey variables and their frequencies.

vi) Number of surveys administered = variable CASEID.

CASEID	**Sequential record number**
Total Cases:	130,880 (Range of valid codes: 1-131535)

vii) Self perceived health of respondents = variable gena_01

gena_01 **Self-perceived health**

Percent	N	Value	Label
22.9	29,953	**1**	EXCELLENT
35.5	46,442	**2**	VERY GOOD
27.5	36,037	**3**	GOOD
10.5	13,715	**4**	FAIR
3.6	4,674	**5**	POOR
	38	**7**	DONT KNOW
	21	**8**	REFUSAL
100.0	**130,880**		**Total**

viii) The barriers to improving health identified are:

- lack of will power (variable ciha_6a)
- lack of time (variable ciha_6b)
- to tired (variable ciha_6c)
- too difficult (variable ciha_6d)
- too costly (variable ciha_6e)
- too stressed (variable ciha_6f)
- disability/health problem (variable ciha_6g)
- other (variable ciha_6h)

The frequencies of the different variables are shown below

ciha_6a **Barrier - lack will power**

Percent	N	Value	Label

38.2	9,210	**1**	YES

61.8	14,888	**2**	NO
	100,540	**6**	NOT APPLICABLE
	15	**7**	DONT KNOW
	2	**8**	REFUSAL
	6,225	**9**	NOT STATED
100.0	**130,880**		**Total**

ciha_6b **Barrier - lack of time**

Percent	N	Value	Label
32.3	7,779	**1**	YES
67.7	16,319	**2**	NO
	100,540	**6**	NOT APPLICABLE
	15	**7**	DONT KNOW
	2	**8**	REFUSAL
	6,225	**9**	NOT STATED
100.0	**130,880**		**Total**

ciha_6c **Barrier - too tired**

Percent	N	Value	Label
3.8	923	**1**	YES
96.2	23,175	**2**	NO
	100,540	**6**	NOT APPLICABLE
	15	**7**	DONT KNOW
	2	**8**	REFUSAL
	6,225	**9**	NOT STATED
100.0	**130,880**		**Total**

ciha_6d **Barrier - too difficult**

Percent	N	Value	Label
3.3	800	**1**	YES
96.7	23,298	**2**	NO
	100,540	**6**	NOT APPLICABLE
	15	**7**	DONT KNOW
	2	**8**	REFUSAL
	6,225	**9**	NOT STATED
100.0	**130,880**		**Total**

ciha_6e **Barrier - too costly**

Percent	N	Value	Label
4.0	964	**1**	YES
96.0	23,134	**2**	NO
	100,540	**6**	NOT APPLICABLE
	15	**7**	DONT KNOW
	2	**8**	REFUSAL
	6,225	**9**	NOT STATED
100.0	**130,880**		**Total**

ciha_6f **Barrier - too stressed**

Percent	N	Value	Label
6.7	1,626	**1**	YES
93.3	22,472	**2**	NO
	100,540	**6**	NOT APPLICABLE
	15	**7**	DONT KNOW
	2	**8**	REFUSAL
	6,225	**9**	NOT STATED
100.0	**130,880**		**Total**

ciha_6g **Barrier - disab./health prob.**

Percent	N	Value	Label
8.4	2,036	**1**	YES
91.6	22,062	**2**	NO
	100,540	**6**	NOT APPLICABLE
	15	**7**	DONT KNOW
	2	**8**	REFUSAL
	6,225	**9**	NOT STATED
100.0	**130,880**		**Total**

ciha_6h **Barrier - other**

Percent	N	Value	Label
11.5	2,781	**1**	YES
88.5	21,317	**2**	NO
	100,540	**6**	NOT APPLICABLE
	15	**7**	DONT KNOW
	2	**8**	REFUSAL
	6,225	**9**	NOT STATED
100.0	**130,880**		**Total**

b) From the UT/DLS: Microdata Analysis and Subsetting page click on "Other subscribers" next to the "Survey of labour and income dynamics (SLID)".

i) Left click on "Data" next to "Survey of labour and income dynamics, wave 11, 2003 **census families** *[Version 2, loaded 2007/02/16]*". Left click on the "Codebook" tab at the top of the page. This will bring you to the survey codebook. Left click on "Sequential Variable List" on the left had side of the page and then select any variable heading and then any individual variable. This will bring you to the list of all the survey variables and their frequencies.

The percentage of families who received social assistance in the reference year is shown with variable **fmsaf46**.

fmsaf46	**Census family rec'd SA (social assistance) in reference year**		
Percent	**N**	**Value**	**Label**
8.2	2,755	**1**	Yes
91.8	30,706	**2**	No
100.0	**33,461**		**Total**

ii) The percentage of census families who received employment insurance is the reference year is shown with variable **fmuif46**.

fmuif46	**Census family rec'd EI (employment insurance) in reference year**		
Percent	**N**	**Value**	**Label**
18.2	6,098	**1**	Yes
81.8	27,363	**2**	No
100.0	**33,461**		**Total**

iii) Left click on "Data" next to "Survey of labour and income dynamics, wave 11, 2003 **economic families** *[Version 3, loaded 2007/02/19]*". Left click on the "Codebook" tab at the top of the page. This will bring you to the survey codebook. Left click on "Sequential Variable List" on the left had side of the page and then select any variable heading and then any individual variable. This will bring you to the list of all the survey variables and their frequencies.

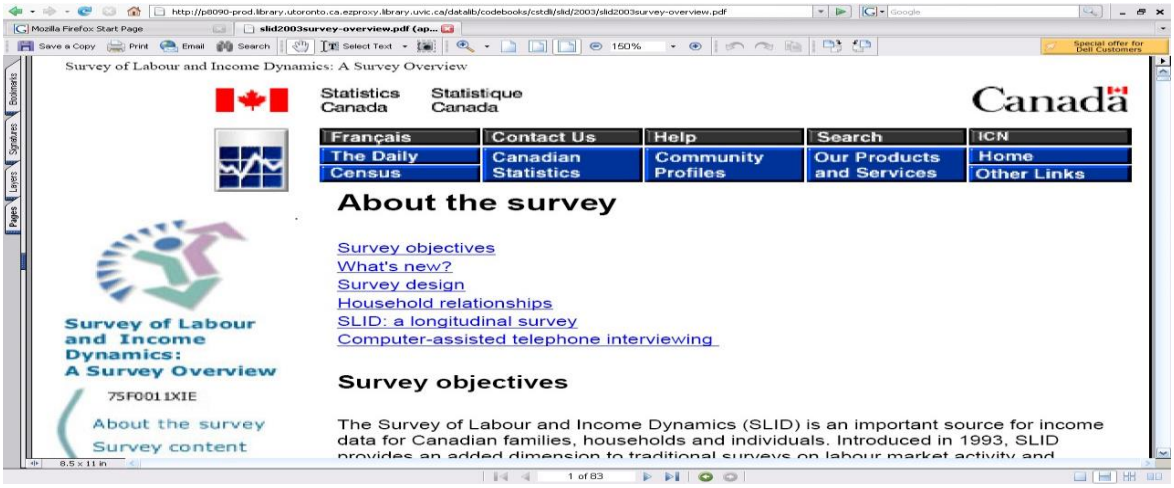
The percentage of economic families who received social assistance (SA) in the reference year is shown with variable **fmsaf27**.

fmsaf27	**Family received SA (social assistance) in reference year**		
Percent	**N**	**Value**	**Label**
8.9	2,650	**1**	Yes
91.1	27,196	**2**	No
100.0	**29,846**		**Total**

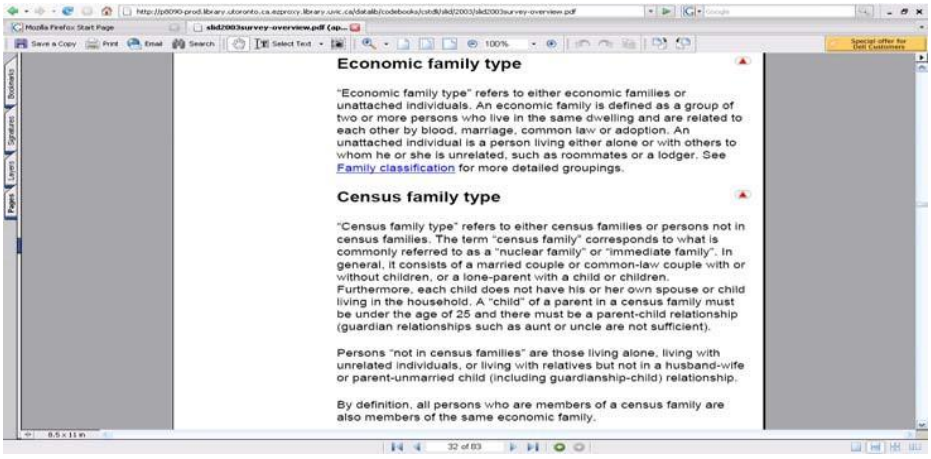
iv) The percentage of economic families who received employment insurance (EI) in the reference year is shown with variable **fmuif27**.

fmuif27	Family received EI (employment insurance) in reference year		
Percent	**N**	**Value**	**Label**
19.9	5,947	**1**	Yes
80.1	23,899	**2**	No
100.0	**29,846**		**Total**

v) In order to answer this question you must search the survey documentation. Left click on "Documentation" next to "Survey of labour and income dynamics, wave 11, 2003". Left click on "Survey overview" under "**File 1. Documentation**". This will take you to a pdf version of a Statistics Canada webpage.



Scroll down the left hand side and click on "Notes and Definitions". Click on "Family" and then scroll down the page until you come to the definitions of economic family type and census family type.



Selecting a Dataset

The SDA tool provides access to over a hundred datasets and students may not know what dataset is suitable for their topic. There are several ways in which students and researchers can narrow their selection, assuming they have a topic already in mind. If you have not selected a topic, the resources listed here will be unable to help you.

1. Learning Objectives

Upon completion of this section you will be able to use various search tools and identify individuals to help you select a dataset for your research.

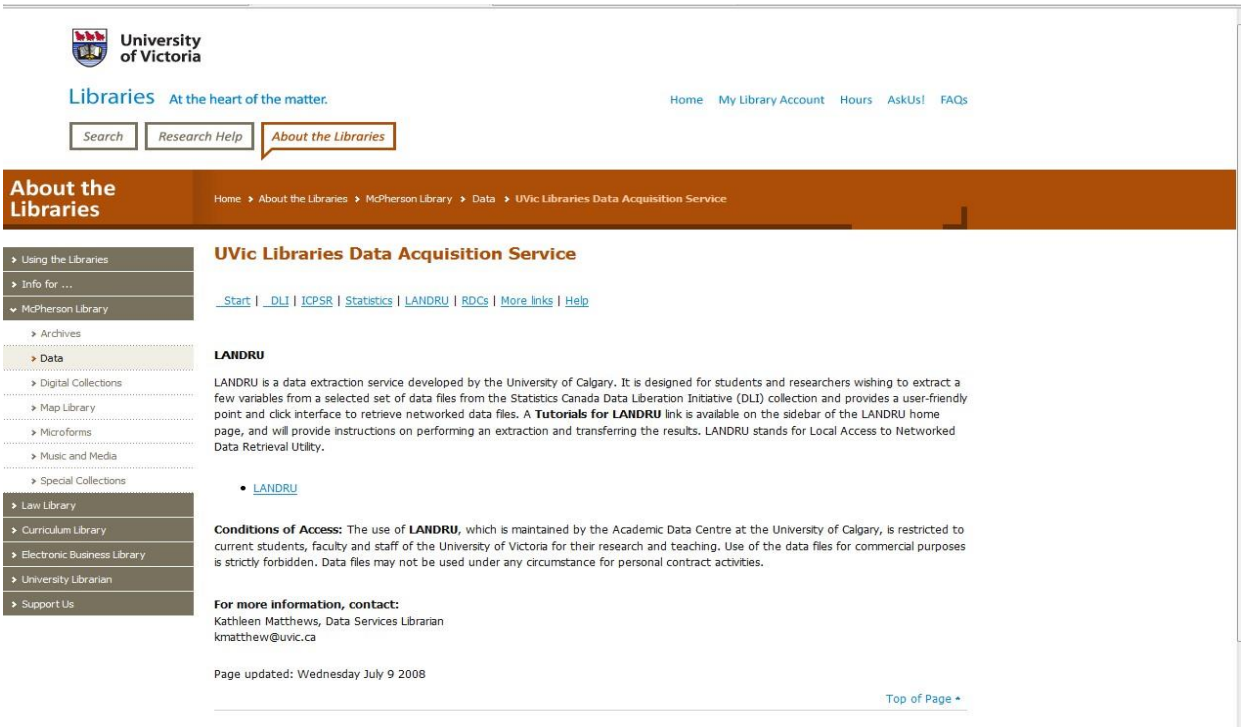
2. UT/DLS

A search tool within the UT/DLS that searches across the included data sets is currently being developed and should be available soon.

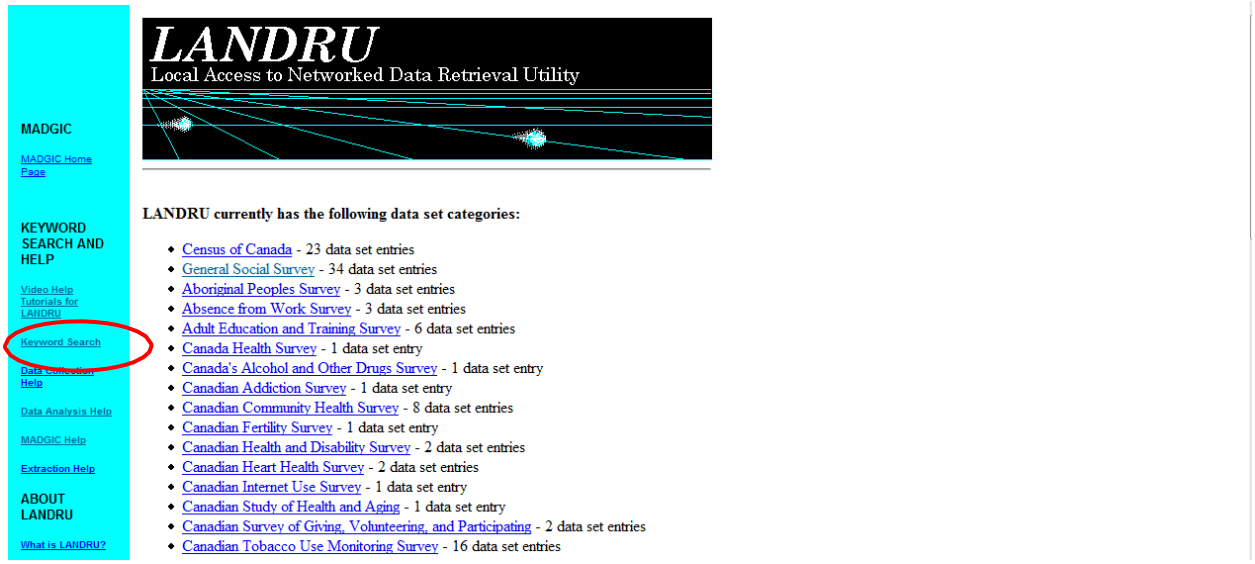
~~3. LANDRU~~ Decomissioned

LANDRU was a data extraction service developed by the University of Calgary. It is designed for students and researchers wishing to extract a few variables from a selected set of data files from the Statistics Canada Data Liberation Initiative (DLI) collection and provides a user-friendly point and click interface to retrieve networked data files. It also includes an easy to use search feature for variables.

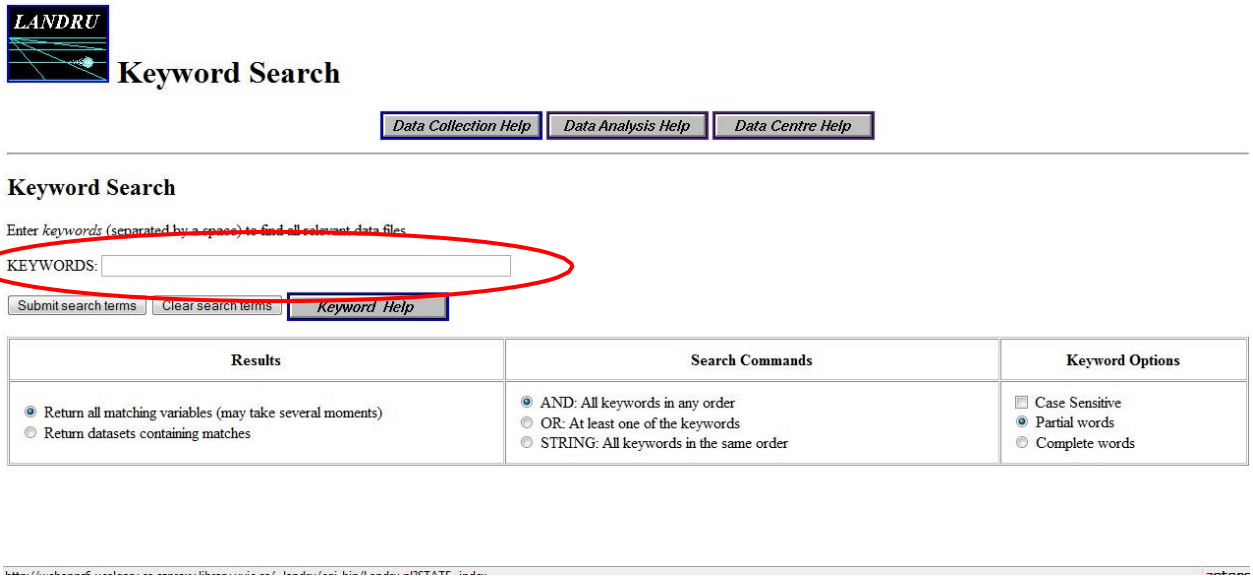
~~To access LANDRU go to~~ http://library.uvic.ca/site/data/landru.html ~~and left click on the LANDRU link.~~



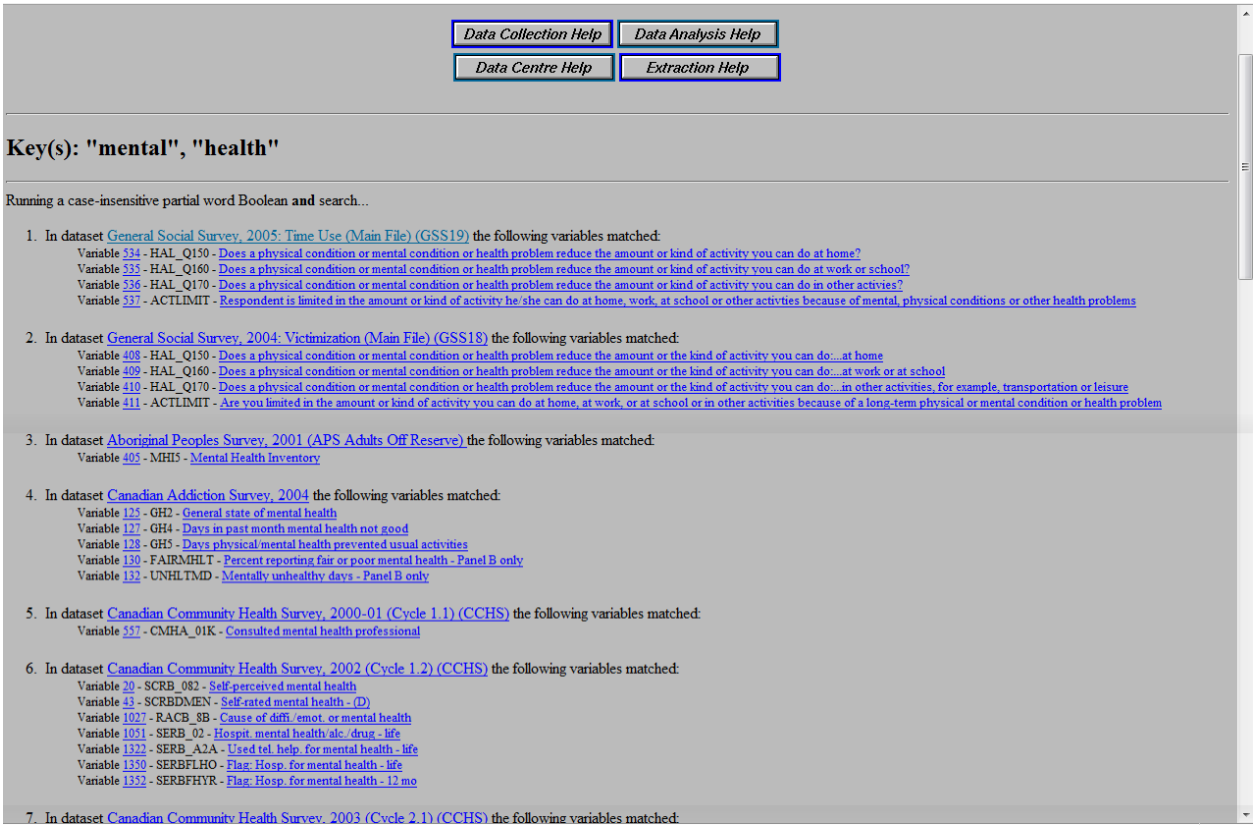
On the left hand side of the page in the green shaded rectangle, is the Keyword Search link. Left click on that link.



Students can then enter in keywords related to their topic in the KEYWORDS: search box.



For example, if you were interested in mental health and wanted to know what survey's asked questions about mental health, you could enter "mental health" in the keyword box and LANDRU would return a list of surveys and the relevant questions related to mental health. You can click on anyone of the questions to see what the response options are along with the frequency distribution of responses. You can use the list that LANDRU provides, perhaps copying it into a word file so that you have a record of the search in your files, to explore the relevant datasets and questions using the SDA tool.



4. Geospatial & Social Sciences Data Librarian

daniel Brendle-Moczuk is the Data Services librarian at the University of Victoria. He is available to answer questions and concerns related to locating data, Statistics Canada, the SDA tool, and all other things data related. Providing you have a topic in mind, daniel can help you find an appropriate dataset.
danielbm@uvic.ca 250-853-3619

5. Instructor

Many faculty are familiar with the data that is available through the UT/DLS tool.
Lindsay Tedds, the coauthor of this tutorial and instructor for the Research Design course in the MPA program at the University of Victoria, is quite familiar with the DLI data sets and many of the other datasets that are accessible through the SDA tool. She regularly uses these datasets in her own research and frequently reads papers by other scholars who have used these datasets. University of Victoria MPA students can seek help selecting a data set during the instructors posted office hours.

Frequency Distributions and Graphs

We now turn to basic **univariate analysis** using the UT/DLS service, particularly basic frequency distributions and graphs.

1. Learning Objectives

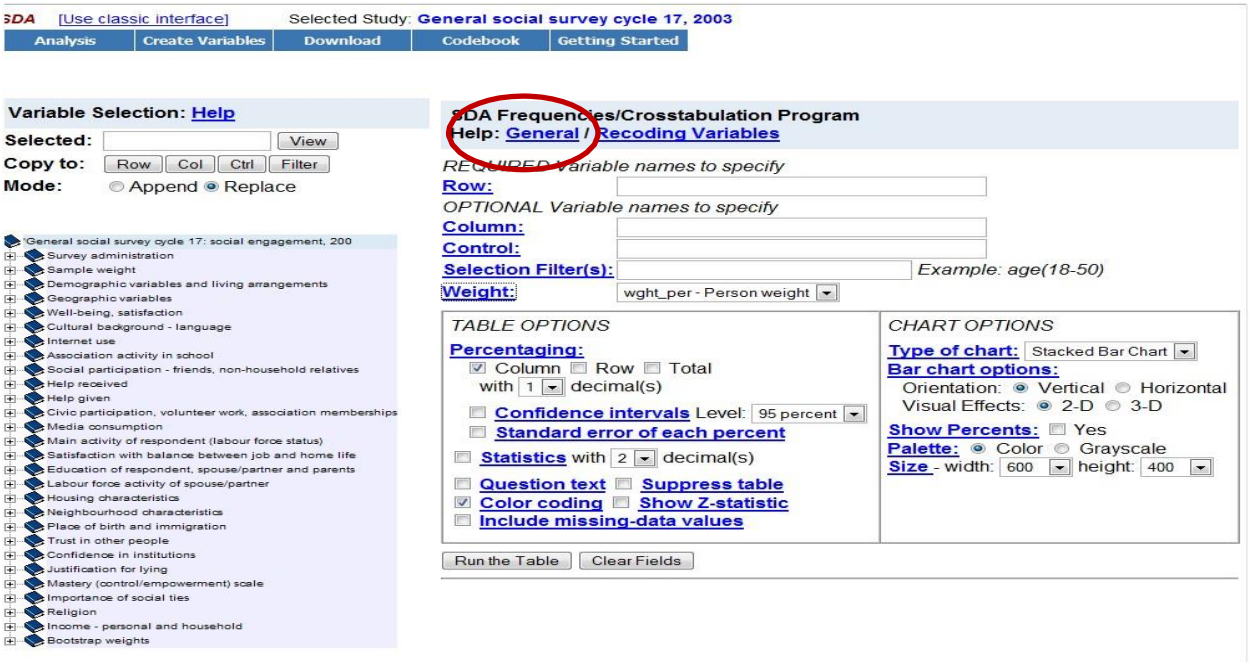
Upon completion of this section you will be able to:

- Calculate variable frequency distributions for entire samples and sample subgroups
- Create bar and pie graphs using the UT/DLS graphing tool

2. Frequency Distributions

A frequency distribution is a count of the number of cases that take on each value of a variable. A univariate frequency distribution table summarizes the categorical information for a single variable, while a cross tabulation (which is covered in the next section) presents categorical info on two variables. Measures of central tendency (mean, mode, median) are based on frequency distributions.

The UT/DLS Frequencies or Cross tabulations analysis program allows us to calculate variable frequency distributions for an entire survey sample or sample subgroups (e.g. female respondents; respondents greater than 65 years of age). To read the description of the program left click on "General" Help under the SDA Frequencies/Crosstabulation Program.



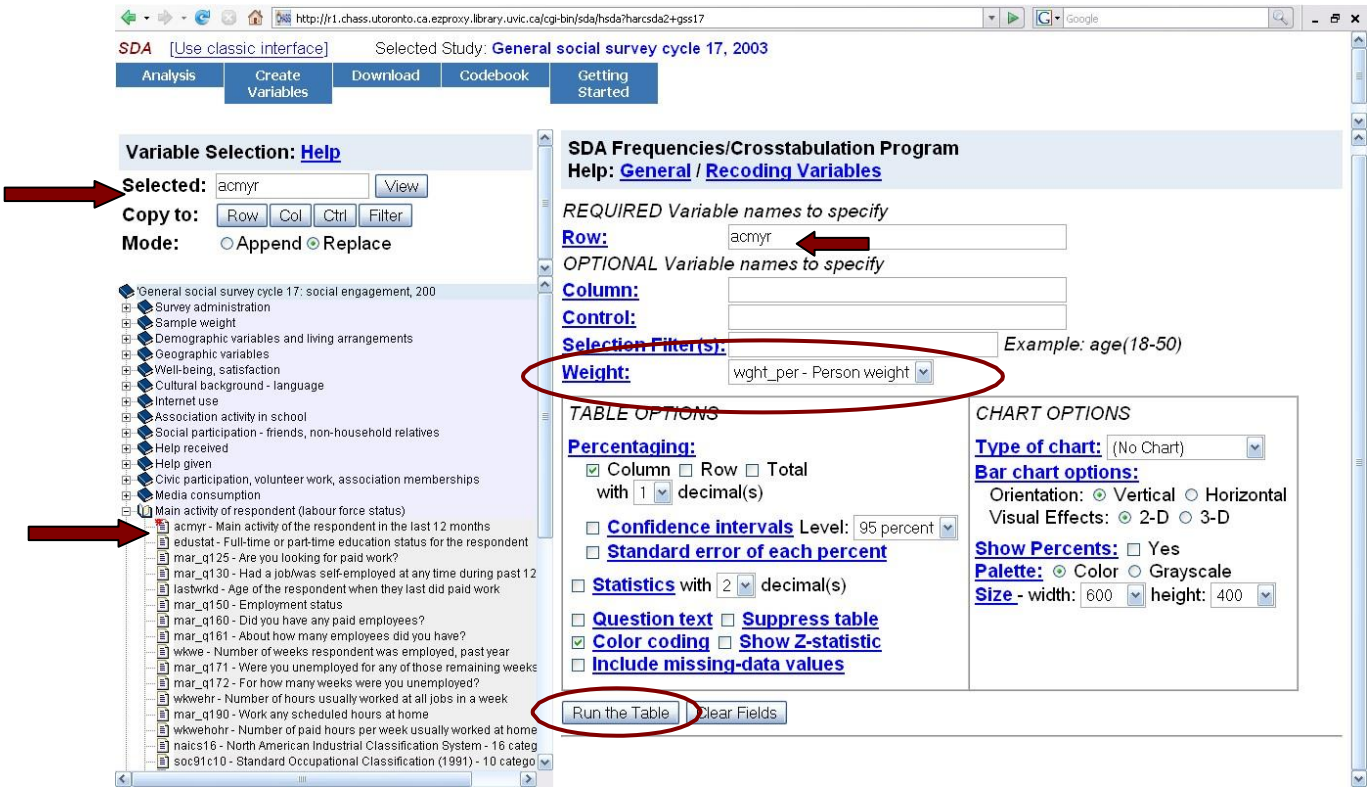
2.1 Entire Sample

We will use the General Social Survey (GSS) on Social Engagement (Cycle 17) to illustrate how to calculate a variable's frequency distribution for an entire survey sample.

- ✓ Left click the "Data" link beside General social survey on social engagement (cycle 17), 2003.

We will calculate the frequency distribution for the variable, labour force status of respondents.

- ✓ In the "Variable Selection" tool, double click on the "Main activity of respondent (labour force status)" variables heading.
- ✓ Double click on the variable "acmyr – Main activity of the respondent in the last 12 months".
- ✓ The variable name will appear in the "Selected" box at the top of the page.
- ✓ Select the "Row" button next to "Copy to" in order to copy the variable name into the UT/DLS Frequencies/Crosstabulation Program.
- ✓ Ensure that in the "Weight" box "wght_per – Person weight" is selected. The default in the UT/DLS service is to produce weighted frequencies. If you do not use the sample weights, then you will be reporting characteristics of the sample. These characteristics cannot be used to infer to the underlying population. *Note*: the frequencies distributions displayed in the Codebook are for the unweighted variables.
- ✓ Click the "Run the Table" button.



The following frequency distribution table will appear:

Variables

Role	Name	Label	Range	MD	Dataset
Row	**acmyr**	Main activity of the respondent in the last 12 months	1-9	98,99	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1

Frequency Distribution

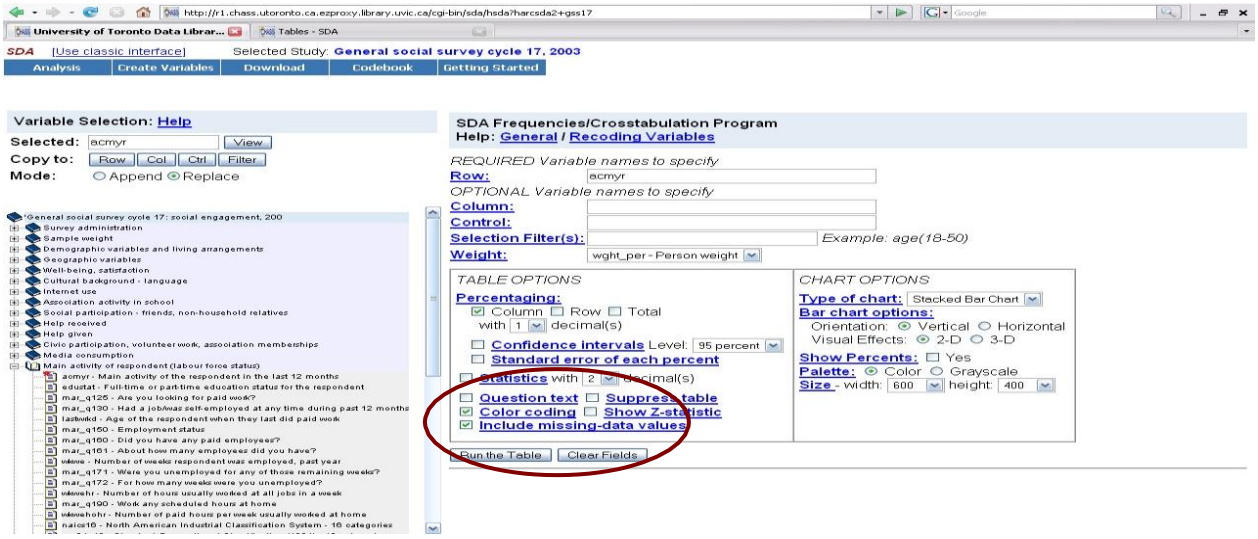
Cells contain: **-Column percent** **-N of cases**		**Distribution**
acmyr	1: Working at a paid job or business	**56.2** 14,212,846
	2: Looking for paid work	**2.3** 590,541
	3: Going to school	**12.3** 3,104,581
	4: Caring for children	**4.5** 1,142,805
	5: Household work	**5.0** 1,257,477
	6: Retired	**16.4** 4,155,527
	7: Maternity / paternity leave	**.2** 58,679
	8: Long term illness	**1.9** 485,026
	9: Other	**1.2** 299,610
	COL TOTAL	***100.0*** *25,307,092*

Each row represents a different category of the variable of interest. Within the column are shown both the number and percentage of weighted respondents associated with the variable values.[4]

Notice that the range of variable values displayed is 1-9. People who did not state their main activity in the last 12 months (code = 98) or who did not know it (code = 99) are omitted from the distribution.

If you want to also display missing data:

- ✓ Ensure that you select "Include missing-data values" in the table options box of the UT/DLS Frequencies/Crosstabulations program prior to running the table.



The following frequency distribution table will appear and you will notice that two categories were added to the table: "98: Not Stated" and "99: Don't know".

Variables

Role	Name	Label	Range	MD	Dataset
Row	**acmyr**	Main activity of the respondent in the last 12 months	1-9	98,99	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1

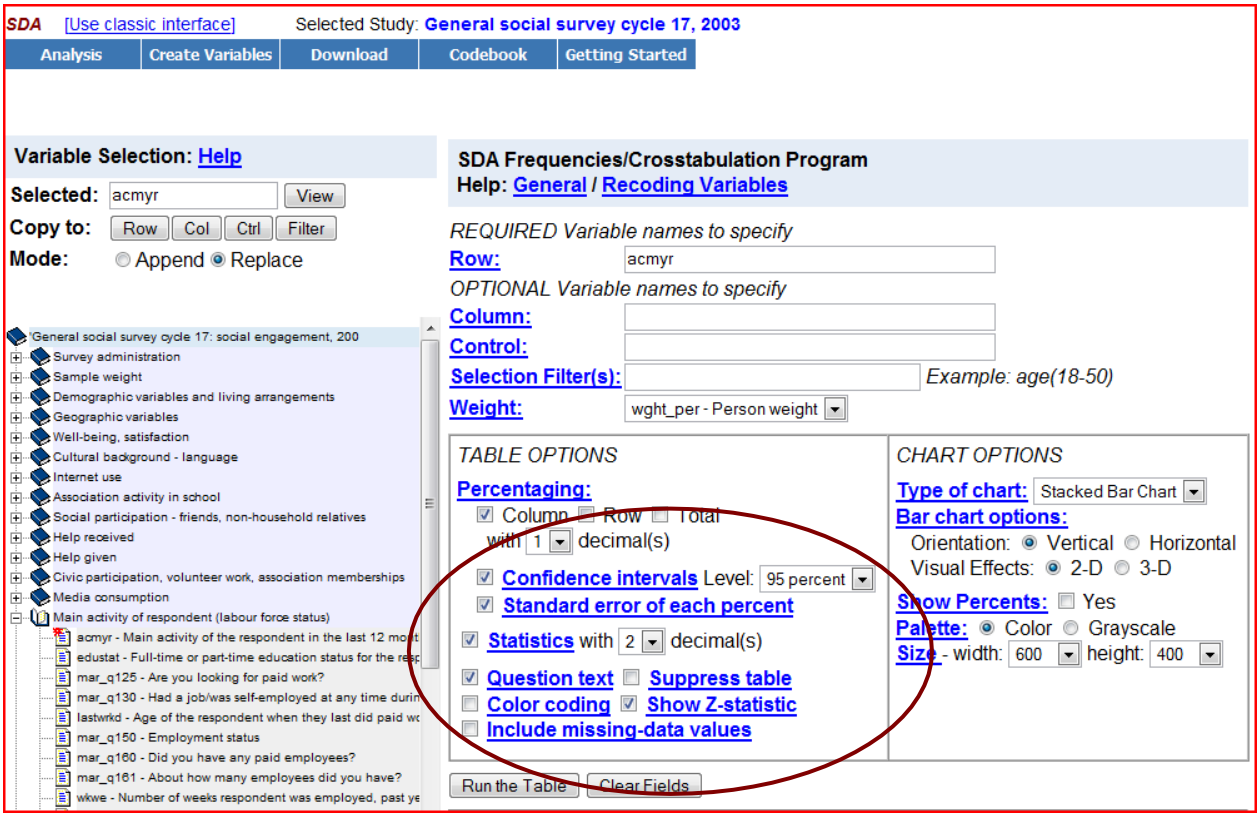
[4] When constructing tables, you need to ensure that enough observations are present in a category to be informative. You must have at least five observations but the typical rule of thumb is 10-15 observations but remember this rule of thumb is based on unweighted frequencies. Therefore, it is recommended that users always double check the unweighted frequencies before finalizing their analysis. If there are not enough observations in one or more cells,users will have to consider collapsing the categories of the variables (e.g. from 5 categories to 3 categories) or censoring their results. Remember, the key is that enough observations have to appear such that you can make inferences about the population and this rule of thumb is attempting to ensure that.

Frequency Distribution		
Cells contain: -Column percent -N of cases		**Distribution**
acmyr	1: Working at a paid job or business	**55.6** 14,212,846
	2: Looking for paid work	**2.3** 590,541
	3: Going to school	**12.2** 3,104,581
	4: Caring for children	**4.5** 1,142,805
	5: Household work	**4.9** 1,257,477
	6: Retired	**16.3** 4,155,527
	7: Maternity / paternity leave	**.2** 58,679
	8: Long term illness	**1.9** 485,026
	9: Other	**1.2** 299,610
	98: Not stated	**.9** 226,486
	99: Don't know	**.0** 10,845
	COL TOTAL	***100.0*** *25,544,423*

You can also produce various statistical measures by selecting a few more options in the table options box of the UT/DLS Frequencies/Crosstabulations program.

- ✓ Ensure that you de-select "Include missing-data values" in the table options box of the UT/DLS Frequencies/Crosstabulations program from the previous activity.

- ✓ In the table options box of the UT/DLS Frequencies/Crosstabulations program select:
  - Confidence Intervals - Level 95%
  - Standard error of each percent
  - Statistics – with 2 decimal(s)
  - Show Z-statistics
- ✓ Click the "Run the Table" button.



With these options selected, you will notice that a 95% confidence interval, standard errors, and the z-statistics have been added into each cell in the frequency distribution and an additional box has been added at the end of the frequency distribution that contains summary statistics.

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Row	**acmyr**	Main activity of the respondent in the last 12 months	1-9	98,99	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1

Frequency Distribution		
Cells contain: -Column percent -Confidence intervals (95 percent) -SRS Std Errs -Z-statistic -N of cases		**Distribution**
acmyr	1: Working at a paid job or business	**56.2** *(56.1-56.2)* .01 .00

		14,212,846
	2: Looking for paid work	**2.3** *(2.3-2.3)* .00 .00 590,541
	3: Going to school	**12.3** *(12.3-12.3)* .01 .00 3,104,581
	4: Caring for children	**4.5** *(4.5-4.5)* .00 .00 1,142,805
	5: Household work	**5.0** *(5.0-5.0)* .00 .00 1,257,477
	6: Retired	**16.4** *(16.4-16.4)* .01 .00 4,155,527
	7: Maternity / paternity leave	**.2** *(0.2-0.2)* .00 .00 58,679
	8: Long term illness	**1.9** *(1.9-1.9)* .00 .00 485,026
	9: Other	**1.2** *(1.2-1.2)* .00 .00 299,610
	COL TOTAL	***100.0*** ---

		--- --- *25,307,092*

Summary Statistics

Mean =	2.67	Std Dev =	2.20	Coef var =	.82
Median =	1.00	Variance =	4.82	Min =	1.00
Mode =	1.00	Skewness =	.97	Max =	9.00
Sum =	67,486,888.69	Kurtosis =	-.35	Range =	8.00

Inference about the mean:

Std Err =	.00	CV(mean) =	.00

Statistics exclude missing-data and out-of-range values.

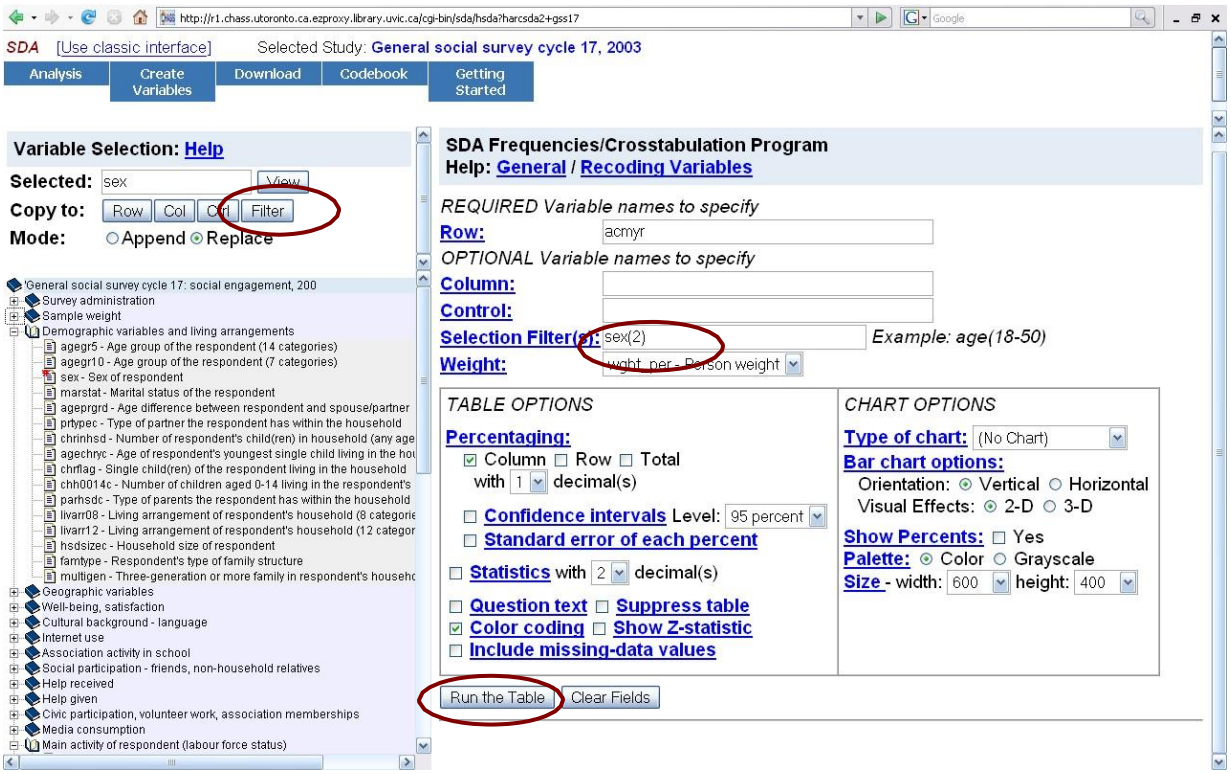
2.2 Subgroups

Often we are only interested in the frequency distribution of a subgroup of a population (e.g. labour force status of female respondents). The UT/DLS allows us to filter survey data to display only the subgroup of interest. Suppose we are only interested in the labour force status of women.

- ✓ Double click on the "Main activity of respondent (labour force status)" variable heading on the left hand side of the page.
- ✓ Double click on the variable "acmyr – Main activity of the respondent in the last 12 months" which should appear in the "Selected" box.
- ✓ Select the "Row" button next to "Copy to" in order to copy the variable name into the UT/DLS Frequencies/Crosstabulation Program.

In order to filter this variable to determine the frequency distribution of females respondents:

- ✓ Double click on the "Demographic variables and living arrangements" variable heading on the left hand side of the page
- ✓ Double click on the variable "sex – Sex of respondent" which should appear in the "Selected" box.
- ✓ Select the "Filter" button next to "Copy to" in order to copy the variable name into the "Selection Filter(s)" box in the SDA Frequencies/Crosstabulation Program. "sex()" should appear in the Selection Filter(s) box.
- ✓ Between the two brackets type "2" the code for female respondents (consult the survey Codebook to determine the variable value to filter)



✓ Left click "Run the Table"

The resulting frequency distribution table for the subgroup females should look like this:

Variables

Role	Name	Label	Range	MD	Dataset
Row	**acmyr**	Main activity of the respondent in the last 12 months	1-9	98,99	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1
Filter	**sex(2)**	Sex of respondent(=Female)	1-2		1

Frequency Distribution

Cells contain: -Column percent -N of cases		**Distribution**
acmyr	1: Working at a paid job or business	**47.7** 6,135,418
	2: Looking for paid work	**1.7** 224,998
	3: Going to school	**12.3**

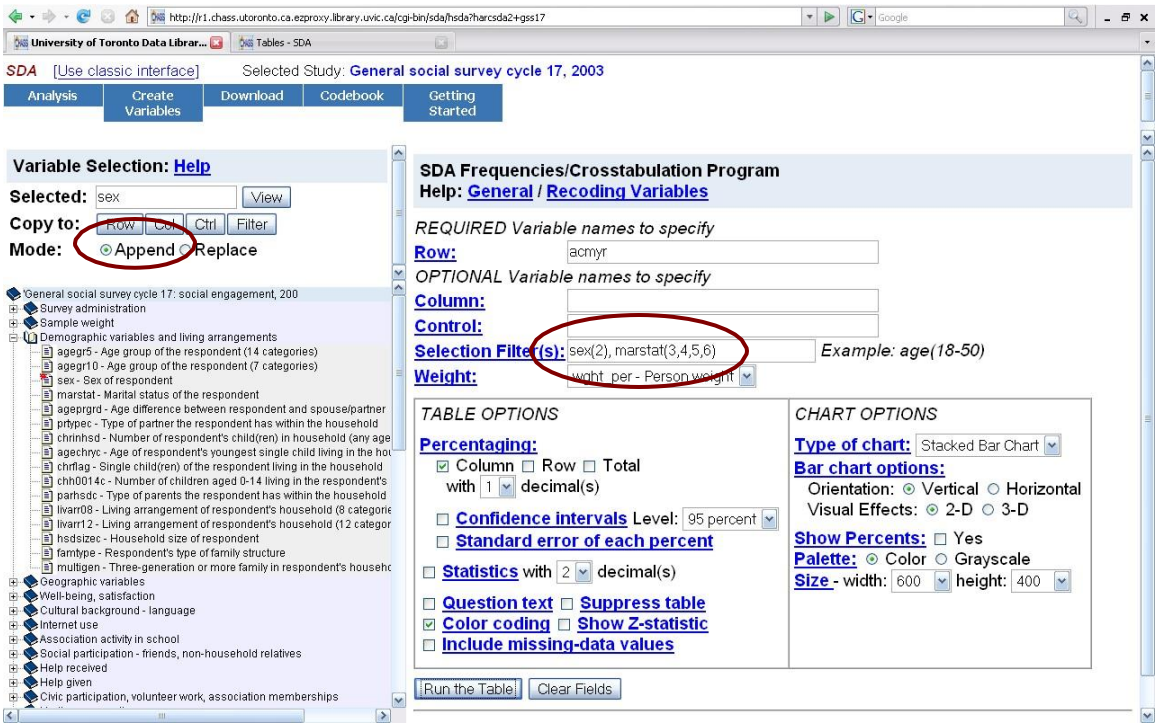
		1,584,305
	4: Caring for children	**8.4** 1,078,932
	5: Household work	**9.0** 1,161,464
	6: Retired	**17.2** 2,207,859
	7: Maternity / paternity leave	**.4** 53,578
	8: Long term illness	**1.9** 244,701
	9: Other	**1.3** 168,804
	COL TOTAL	***100.0*** *12,860,060*

Notice that the frequency distribution varies from that of the entire sample. For example, a smaller percentage of females are working in a paid job or business (47.7) compared to the percentage of the entire population (56.2).

Suppose we are interested in comparing the labor force status frequency distribution for women with and without a partner. We can filter the data based on multiple criteria in order to calculate the relevant frequency distributions:

Women without a partner

- ✓ Use the same Row variable (acmyr) and Selection Filter criteria (sex(2))from the previous example.
- ✓ In the Variable Selection program double click on the "Demographic variables and living arrangements" variable.
- ✓ Double click the variable "marstat – Marital status of the respondent" which should appear in the "Selected" box.
- ✓ Select the "Append" option beside "Mode" and then click the "Filter" button next to "Copy to" in order to copy the variable name into the "Selection Filter(s)" box. ***Note***: if "Replace" is selected it will erase any variables already placed within the selection filter criteria, in this case, the variable "sex".
- ✓ Consult the Codebook to determine the variable values for women without a partner (widowed = 3, separated = 4, divorced = 5, single (never married) = 6).
- ✓ In the "marstat" brackets enter "3,4,5,6" or "3-6" to indicate the variable values we want filtered.
- ✓ In the "Weight" box ensure that "wght-per = Person weight" is selected.



✓ Left click "Run the Table"

Variables

Role	Name	Label	Range	MD	Dataset
Row	**acmyr**	Main activity of the respondent in the last 12 months	1-9	98,99	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1
Filter	**sex(2)**	Sex of respondent(=Female)	1-2		1
Filter	**marstat(3,4,5,6)**	Marital status of the respondent	1-6	8,9	1

Frequency Distribution

Cells contain: -Column percent -N of cases		**Distribution**
acmyr	1: Working at a paid job or business	**40.0** 2,110,745
	2: Looking for paid work	**2.0** 103,277
	3: Going to school	**26.6** 1,405,864
	4: Caring for children	**3.0** 156,667

	5: Household work	**5.2** 274,504
	6: Retired	**19.6** 1,032,976
	7: Maternity / paternity leave	**.0** 1,760
	8: Long term illness	**2.2** 114,437
	9: Other	**1.5** 81,577
	COL TOTAL	***100.0*** *5,281,808*

Women with a partner

- ✓ Consult the Codebook to determine the variable values for women with partners (married = 1, living common-law = 2)
- ✓ In the 'marstat' brackets erase the numbers from the previous table run and enter "1,2" or "1-2" to indicate the new variable values we want filtered
- ✓ In the "Weight" box ensure that "wght-per = Person weight" is selected
- ✓ Left click "Run the table"

Variables

Role	Name	Label	Range	MD	Dataset
Row	**acmyr**	Main activity of the respondent in the last 12 months	1-9	98,99	1
Weight	**wght_per**	Person weight	34.2999- 5,234.9148		1
Filter	**sex(2)**	Sex of respondent(=Female)	1-2		1
Filter	**marstat(1,2)**	Marital status of the respondent	1-6	8,9	1

Frequency Distribution

Cells contain: -Column percent -N of cases		**Distribution**
acmyr	1: Working at a paid job or business	**53.2** 4,015,546
	2: Looking for paid work	**1.6** 117,259
	3: Going to school	**2.3** 176,077
	4: Caring for children	**12.2**

		921,826
	5: Household work	**11.7** 886,961
	6: Retired	**15.5** 1,167,124
	7: Maternity / paternity leave	**.7** 51,818
	8: Long term illness	**1.7** 129,131
	9: Other	**1.2** 87,227
	COL TOTAL	***100.0*** *7,552,968*

Notice the differences in the frequency distribution of the two sub-groups. Women with partners are more likely than women without partners to be working at a paid job or business (53.2% vs. 40.0), or be caring for children (12.2% vs. 3.0%), but are less likely to be going to school (2.3% vs. 26.6%), or be retired (15.5% vs. 19.6%).

3. Graphing Using the UT/DLS Tool

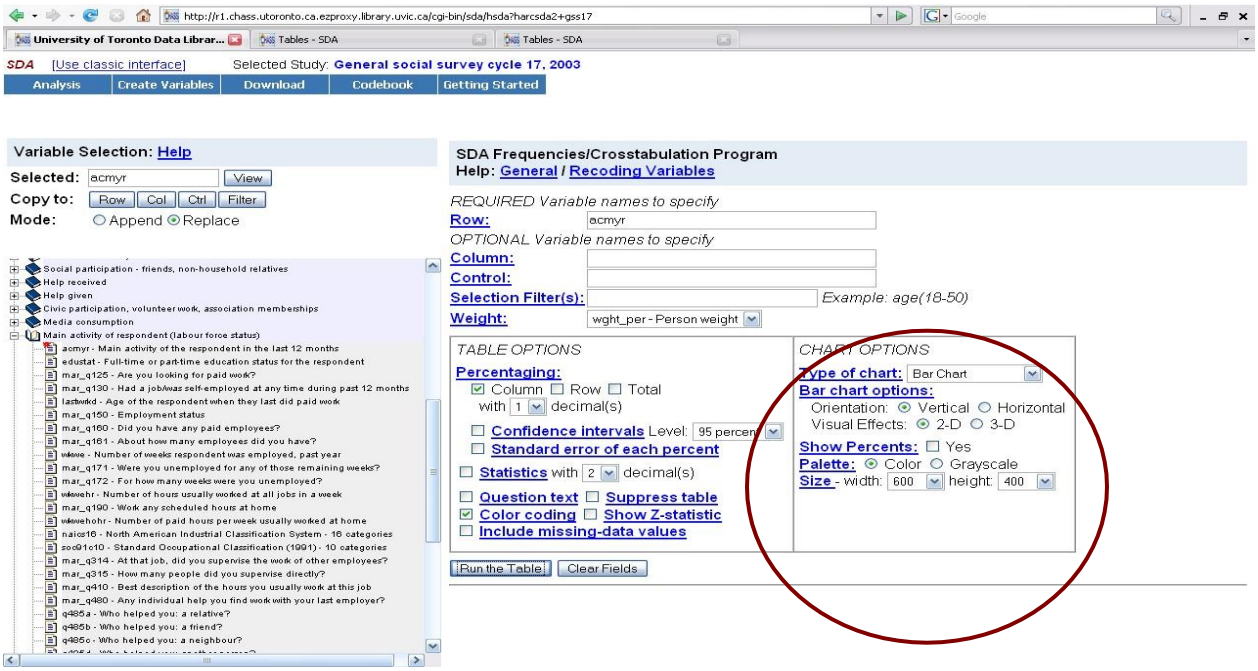
The UT/DLS allows us to also display frequency data graphically using the "CHART OPTIONS" section of the SDA Frequencies/Crosstabulations program. Chart options available include: bar charts, pie charts, and line charts.

3.1 Bar Charts

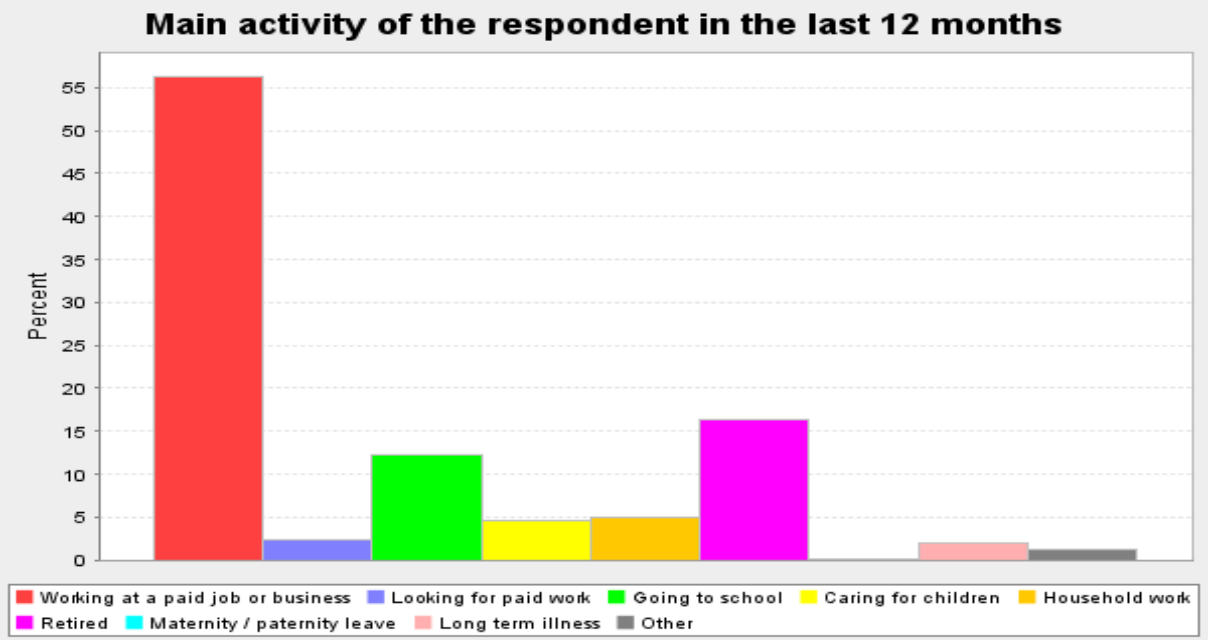
Bar charts are often used to describe categorical data and therefore are commonly used to graphically depict frequency distributions.

Consider our example from section 2.2, the frequency distribution of the variable "acmyr – Main activity of the respondent in the last 12 months". In order to display the variable frequency graphically using a bar chart:

- ✓ Double click on the "Main activity of respondent (labour force status)" variable heading on the left hand side of the page.
- ✓ Double click on the variable "acmyr – Main activity of the respondent in the last 12 months" which should appear in the "Selected" box.
- ✓ Select the "Row" button next to "Copy to" in order to copy the variable name into the UT/DLS Frequencies/Crosstabulation Program.
- ✓ In the *CHART OPTIONS,* in the "Type of Chart" box, select "Bar Chart".
- ✓ Under "Bar Chart Options" select "Vertical" orientation.
- ✓ Left click the "Run the Table" button.



Scroll down past the frequency distribution table to the following bar chart:



To copy the chart into a word document:

- ✓ Right click on the image.
- ✓ Select Copy Image.
- ✓ Place the cursor where you would like the chart to be positioned within your document
- ✓ Select Edit on the top toolbar and click on Paste **or** Type 'Ctrl V'.

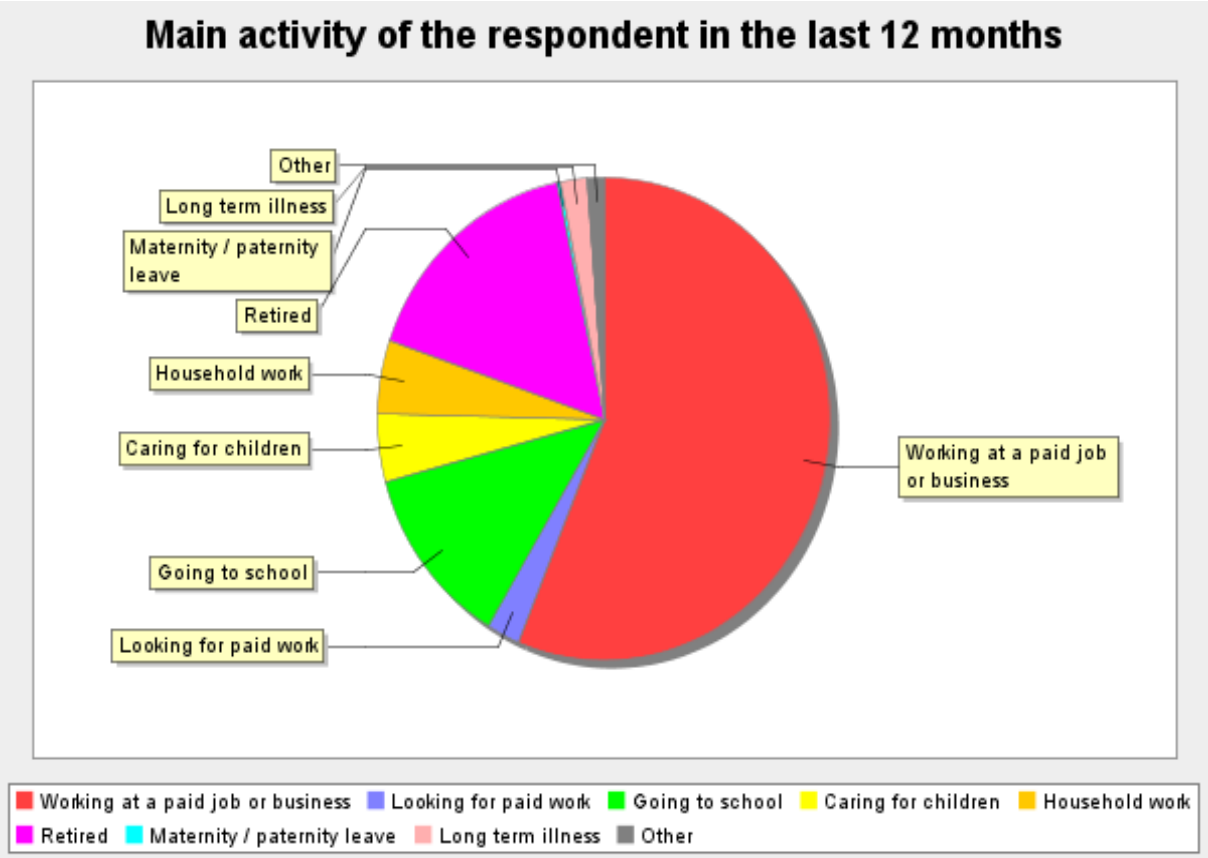
Play around with the other available *CHART OPTIONS* to determine how the graph will change if you select Stacked Bar Chart, Horizontal Orientation, 3-D Visual Effects, Show Percents, and/or Grayscale Palette.

3.2 Pie Charts

Pie charts emphasize the proportion of the number of total respondents in each variable category. The circle represents the total number of people in the sample or sub-group and each segment corresponds to category's share of the total. Segment size is proportional to category frequency.

In order to display the frequency of the variable "acmyr – Main activity of the respondent in the last 12 months" in a pie chart form:

- ✓ In the *CHART OPTIONS*, in the "Type of Chart" box, select "Pie Chart".
- ✓ Click the "Run the Table" button.



4. Conclusion

This tutorial covers the basic material for univariate analysis using the UT/DLS service. In the next section, we will cover bivariate analysis. To ensure that you understand the basic material presented in this section, please work through the following exercises.

5. Exercises

a) Using the GSS Cycle 17, calculate and graph the frequency distribution of the variable "net_q110 – In the past 12 months, did you use the Internet?", found under the "Internet Use" variable heading, for:

i) The entire variable (vertical bar graph)
ii) The subgroup of women (horizontal bar graph in grayscale)
iii) The subgroups of people living in urban and rural areas
iv) The subgroups of people over and under the age of 30

6. Answers

a) i)

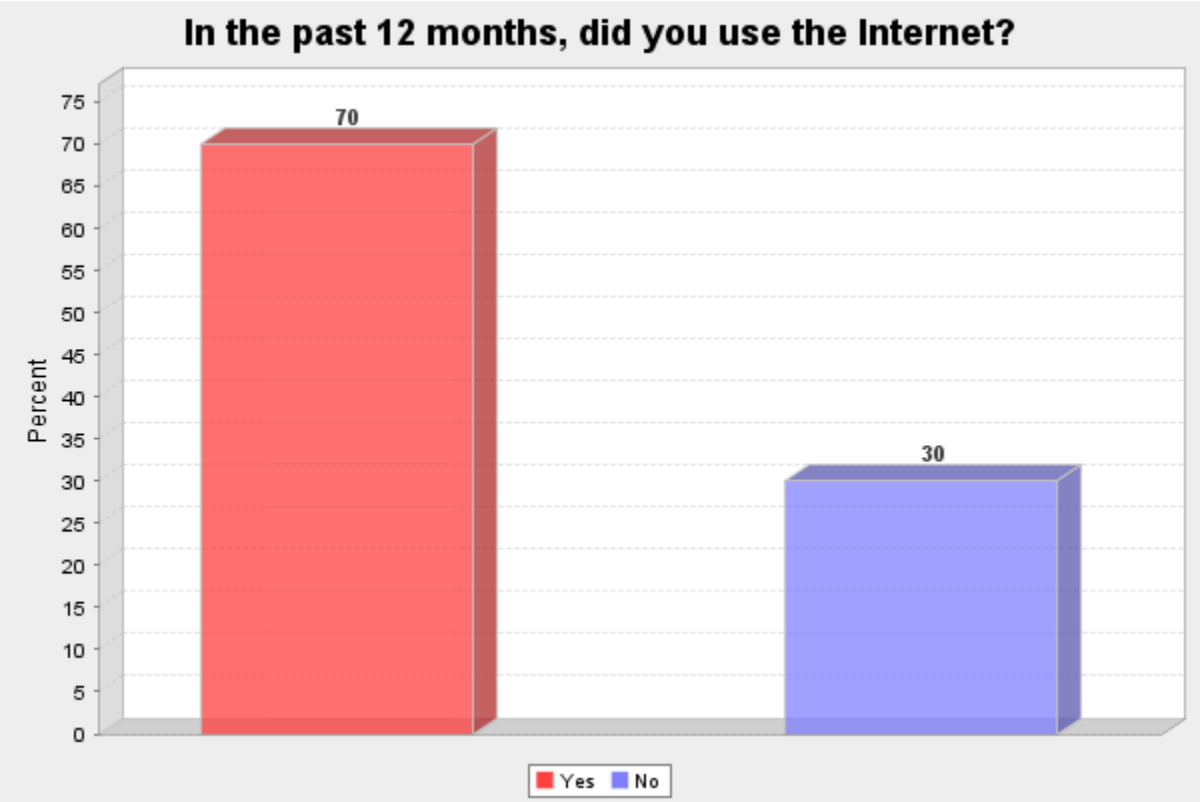
- ✓ Select the variable "net_q110 – In the past 12 months, did you use the Internet?" from the list of variable.
- ✓ Click the "Row" button next to "Copy to".
- ✓ In *CHART OPTIONS* select "Bar Chart".
- ✓ Click "Run the Table"

Variables

Role	Name	Label	Range	MD	Dataset
Row	**net_q110**	In the past 12 months, did you use the Internet?	1-2	8,9	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1

Frequency Distribution

Cells contain: -Column percent -N of cases		**Distribution**
net_q110	1: Yes	**70.0** 17,868,576
	2: No	**30.0** 7,655,378
	COL TOTAL	***100.0*** *25,523,954*

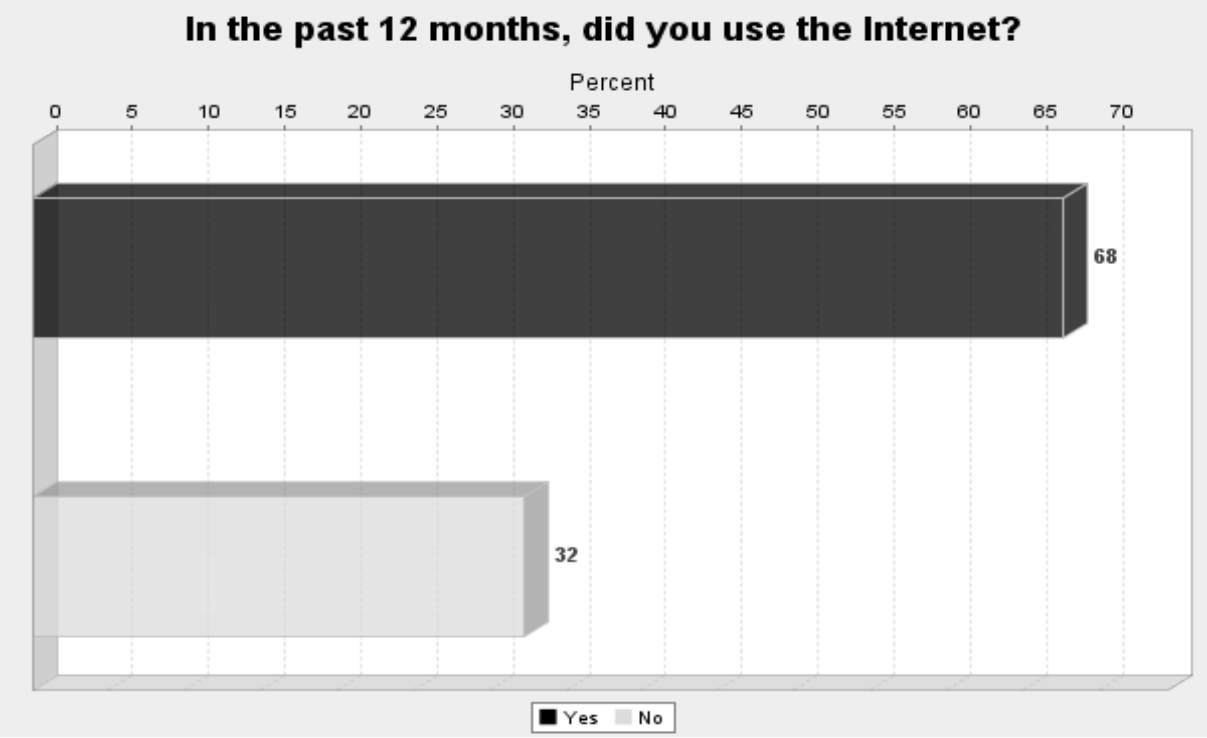


ii)

- ✓ Select the variable "net_q110 – In the past 12 months, did you use the Internet?" from the list of variables.
- ✓ Click the "Row" button next to "Copy to".
- ✓ Double click the variable heading "Demographic variables and living arrangements"
- ✓ Select the variable "sex – Sex of respondent"
- ✓ Select the "Filter" button next to "Copy to" in order to copy the variable name into the "Selection Filter(s)" box in the UT/DLS Frequencies/Crosstabulation Program. "sex()" should appear in the Selection Filter(s) box.
- ✓ Between the two brackets type "2" the code for female respondents (consult the survey Codebook to determine the variable value to filter).
- ✓ In *CHART OPTIONS* select "Bar Chart".
- ✓ Select "Horizontal" Orientation and "Grayscale" Palette.
- ✓ Click "Run the Table"

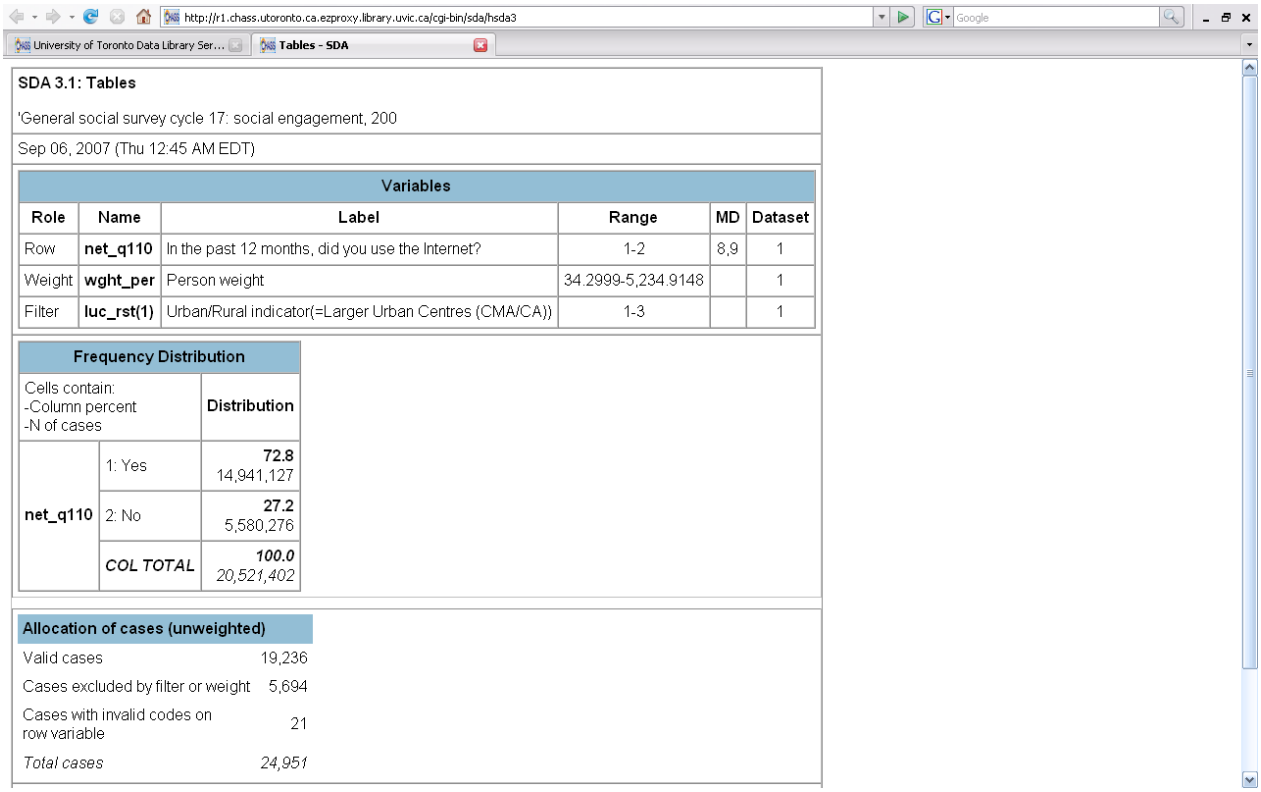
Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Row	**net_q110**	In the past 12 months, did you use the Internet?	1-2	8,9	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1
Filter	**sex(2)**	Sex of respondent(=Female)	1-2		1

Frequency Distribution		
Cells contain: -Column percent -N of cases		**Distribution**
net_q110	1: Yes	**67.7** 8,784,739
	2: No	**32.3** 4,186,139
	COL TOTAL	***100.0*** *12,970,878*



iii)

- ✓ Select the variable “net_q110 – In the past 12 months, did you use the Internet?” from the list of variables.
- ✓ Click the “Row” button next to “Copy to”.
- ✓ Double click the variable heading “Geographic variable”
- ✓ Select the variable “luc_rst – Urban/Rural indicator”
- ✓ Select the “Filter” button next to “Copy to” in order to copy the variable name into the “Selection Filter(s)” box in the UT/DLS Frequencies/Crosstabulation Program. “luc_rst()” should appear in the Selection Filter(s) box.
- ✓ Between the two brackets type “1” the code for large urban centres (consult the survey Codebook to determine the variable value to filter)
- ✓ Click “Run the Table”



http://r1.chass.utoronto.ca.ezproxy.library.uvic.ca/cgi-bin/sda/hsda3

University of Toronto Data Library Ser... | Tables - SDA

SDA 3.1: Tables

'General social survey cycle 17: social engagement, 200

Sep 06, 2007 (Thu 12:45 AM EDT)

Variables

Role	Name	Label	Range	MD	Dataset
Row	**net_q110**	In the past 12 months, did you use the Internet?	1-2	8,9	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1
Filter	**luc_rst(1)**	Urban/Rural indicator(=Larger Urban Centres (CMA/CA))	1-3		1

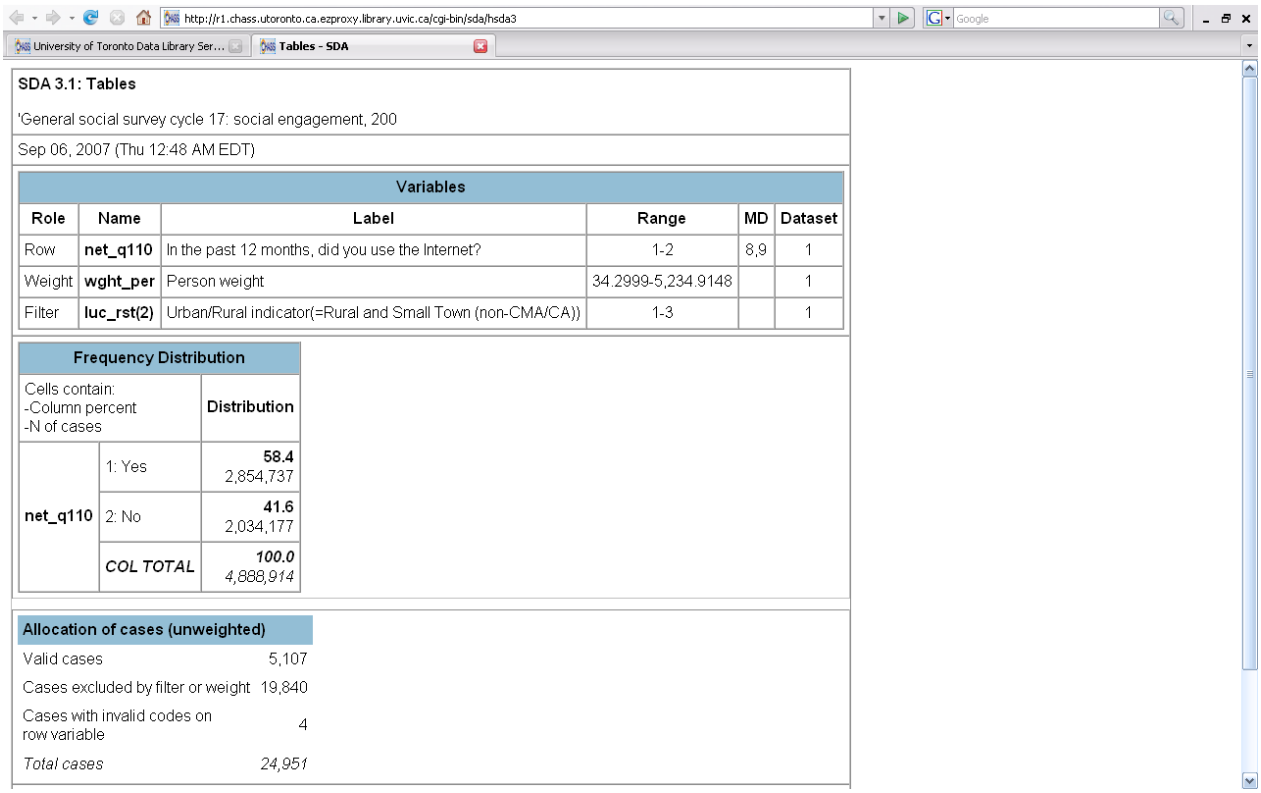
Frequency Distribution

Cells contain: -Column percent -N of cases		**Distribution**
net_q110	1: Yes	**72.8** 14,941,127
	2: No	**27.2** 5,580,276
	COL TOTAL	***100.0*** *20,521,402*

Allocation of cases (unweighted)

Valid cases	19,236
Cases excluded by filter or weight	5,694
Cases with invalid codes on row variable	21
Total cases	*24,951*

- ✓ Change the Selection Filter variable value to “2” the code for Rural and Small Town.
- ✓ Click “Run the Table”



SDA 3.1: Tables

'General social survey cycle 17: social engagement, 200

Sep 06, 2007 (Thu 12:48 AM EDT)

Variables

Role	Name	Label	Range	MD	Dataset
Row	**net_q110**	In the past 12 months, did you use the Internet?	1-2	8,9	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1
Filter	**luc_rst(2)**	Urban/Rural indicator(=Rural and Small Town (non-CMA/CA))	1-3		1

Frequency Distribution

Cells contain: -Column percent -N of cases		**Distribution**
net_q110	1: Yes	**58.4** 2,854,737
	2: No	**41.6** 2,034,177
	COL TOTAL	***100.0*** *4,888,914*

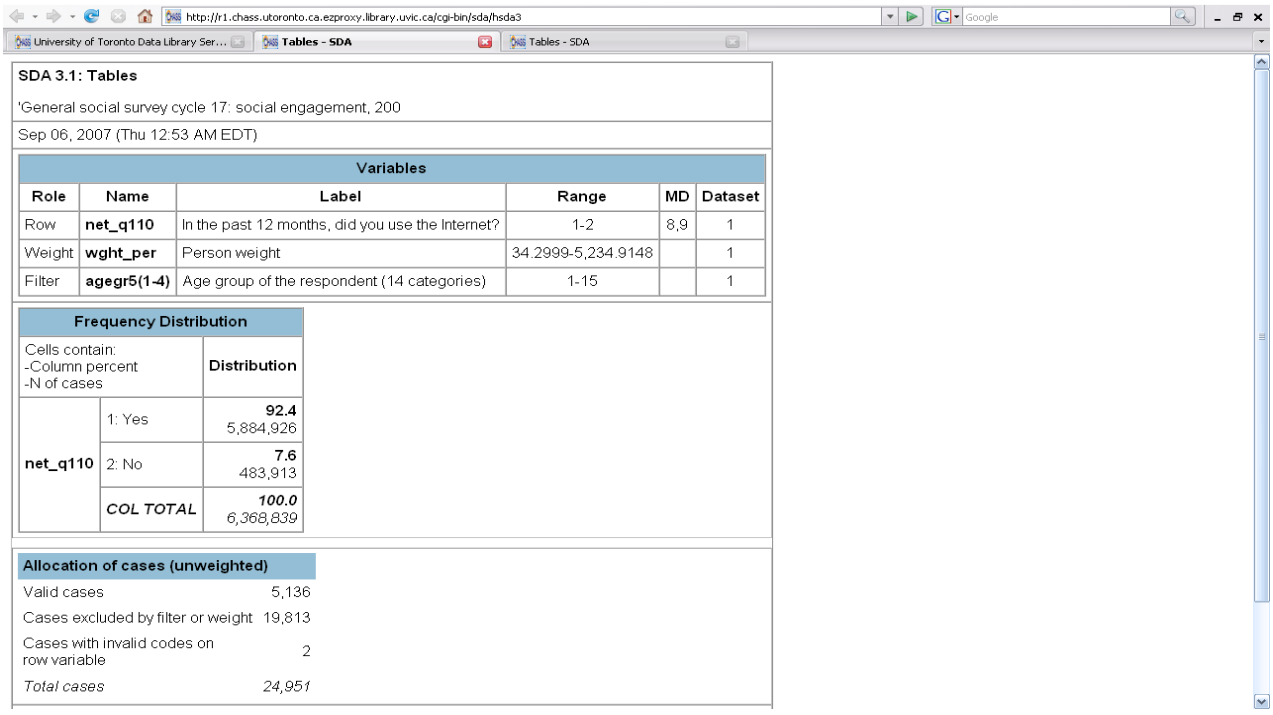
Allocation of cases (unweighted)

Valid cases	5,107
Cases excluded by filter or weight	19,840
Cases with invalid codes on row variable	4
Total cases	*24,951*

Note the differences in the frequency distribution between urban and rural areas. In urban areas 72.8% of respondents reported using the Internet in the past 12 months, compared to only 58.4% of rural and small town respondents.

iv)

- ✓ Select the variable "net_q110 – In the past 12 months, did you use the Internet?" from the list of variables.
- ✓ Click the "Row" button next to "Copy to".
- ✓ Double click the variable heading "Geographic variable"
- ✓ Select the variable "agegr5 – Age group of the respondent (14 categories)"
- ✓ Select the "Filter" button next to "Copy to" in order to copy the variable name into the "Selection Filter(s)" box in the UT/DLS Frequencies/Crosstabulation Program. "agegr5()" should appear in the Selection Filter(s) box.
- ✓ Between the two brackets type "1-4" the code ranges for people under the age of 30 (consult the survey Codebook to determine the variable value to filter)
- ✓ Click "Run the Table"



SDA 3.1: Tables

'General social survey cycle 17: social engagement, 200

Sep 06, 2007 (Thu 12:53 AM EDT)

Variables

Role	Name	Label	Range	MD	Dataset
Row	**net_q110**	In the past 12 months, did you use the Internet?	1-2	8,9	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1
Filter	**agegr5(1-4)**	Age group of the respondent (14 categories)	1-15		1

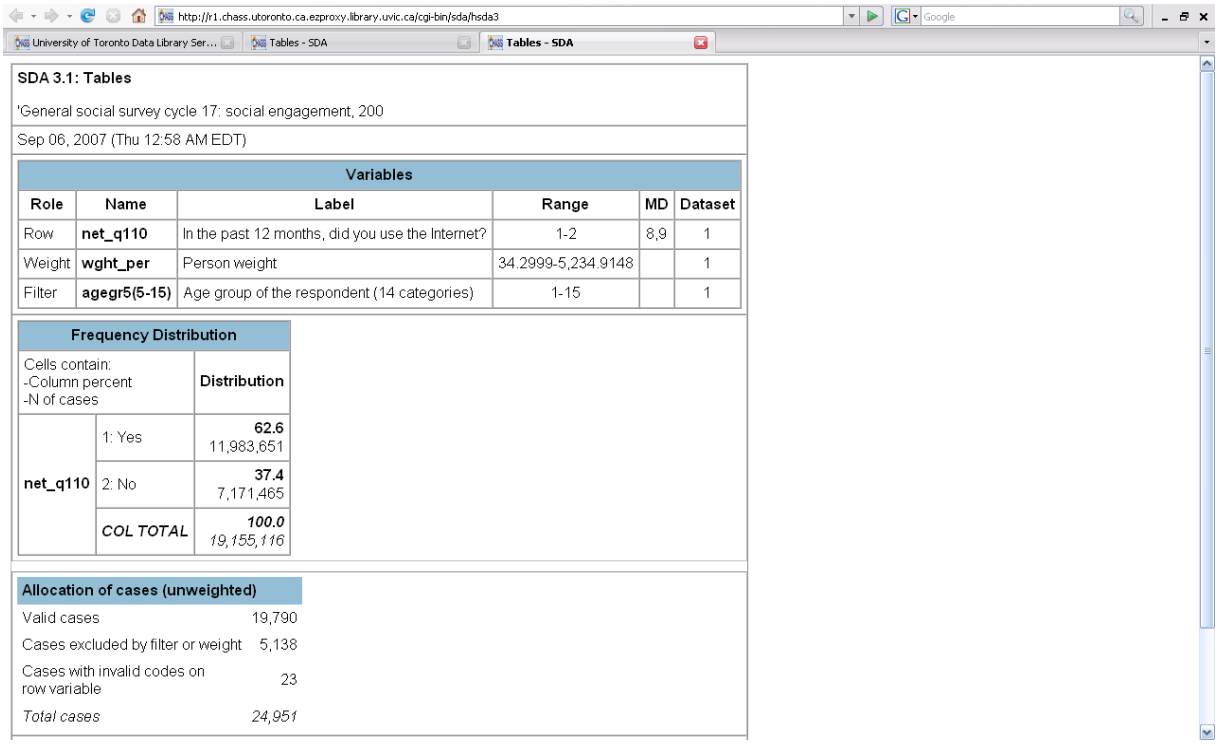
Frequency Distribution

Cells contain: -Column percent -N of cases		**Distribution**
net_q110	1: Yes	**92.4** 5,884,926
	2: No	**7.6** 483,913
	COL TOTAL	***100.0*** *6,368,839*

Allocation of cases (unweighted)

Valid cases	5,136
Cases excluded by filter or weight	19,813
Cases with invalid codes on row variable	2
Total cases	*24,951*

- ✓ Change the Selection Filter variable values to "5-15" the code ranges for persons 30 years of age or over.
- ✓ Click "Run the Table"



SDA 3.1: Tables

'General social survey cycle 17: social engagement, 200

Sep 06, 2007 (Thu 12:58 AM EDT)

Variables

Role	Name	Label	Range	MD	Dataset
Row	**net_q110**	In the past 12 months, did you use the Internet?	1-2	8,9	1
Weight	**wght_per**	Person weight	34.2999-5,234.9148		1
Filter	**agegr5(5-15)**	Age group of the respondent (14 categories)	1-15		1

Frequency Distribution

Cells contain: -Column percent -N of cases		**Distribution**
net_q110	1: Yes	**62.6** 11,983,651
	2: No	**37.4** 7,171,465
	COL TOTAL	***100.0*** *19,155,116*

Allocation of cases (unweighted)

Valid cases	19,790
Cases excluded by filter or weight	5,138
Cases with invalid codes on row variable	23
Total cases	*24,951*

Notice the difference in the frequency distribution between the two subgroups. 92.4% of respondents under 30 reported using the Internet in the past 12 months compared to only 62.6% of persons 30 years of age and older.

Analyzing Bivariate Relationships

1. Learning Objectives

In social science research, we are generally more interested in determining relationships between two or more variables (bivariate and multivariate relationships) then in describing distributions of single variables. This tutorial focuses on different methods used to analyze relationships between two variables. Upon completion of this tutorial you will be able to:

- Perform cross tabulations using the UT/DLS
- Graph cross tabulations
- Calculate statistics using the SDA Frequencies/Crosstabulation Program
- Perform comparison of means calculations
- Calculate confidence intervals
- Compute correlations and comparison of correlations
- Perform simple regression analysis

2. Cross tabulation

Cross tabulation (cross tab) summarizes the relationship(s) between two or more nominal or ordinal variables in tabular format. Cross tabs differ from simple tables in that they "are based more directly upon hypotheses and are structured so as to facilitate an examination of the relationships between variables" (Manheim et al., pp.250).

2.1 Creating Cross tabs

The SDA Frequencies/Crosstabulation program allows you to generate the cross tabulation of two variables.

Let us examine the relationship between age and self-rated health using the Canadian Community Health Survey (CCHS) cycle 3.1. We will organize the cross tab so as to examine the hypothesis that self-rated health declines with age.

- ✓ Left click on "Data" beside Cycle 3.1, 2005 common & optional content of the CCHS.
- ✓ Identify the codes for the variables of interest, self-rated health and age, using the **Variable Selection** tool on the left hand side of the page and then copy the variables into the **UT/DLS Frequencies/Cross tabulations Program** on the right hand side of the page:

Self-rated health variable:

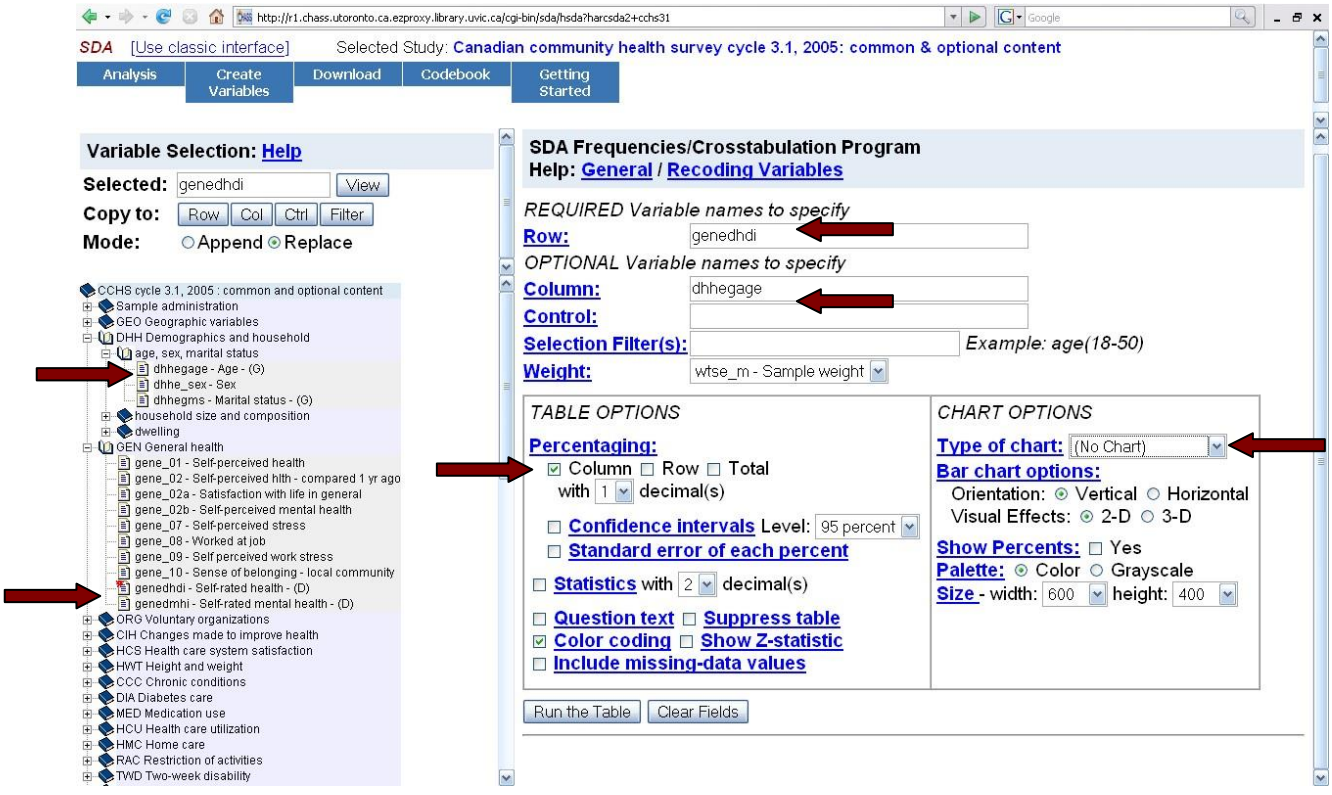
- Double click on the "GEN General health" variables heading.
- Double click on the variable "genedhdi – Self-rated health – (D)" The variable name will appear in the **Selected** box above. **Note:** This variable is derived from the variable "gene 01 – Self-perceived health". For information on how the variable was derived consult the "Derived and Group Variables" section of the survey's documentation

- Click on the "Row" button. The variable name will then appear in the **Row** box of the **SDA Frequencies/Cross tabulations Program.**

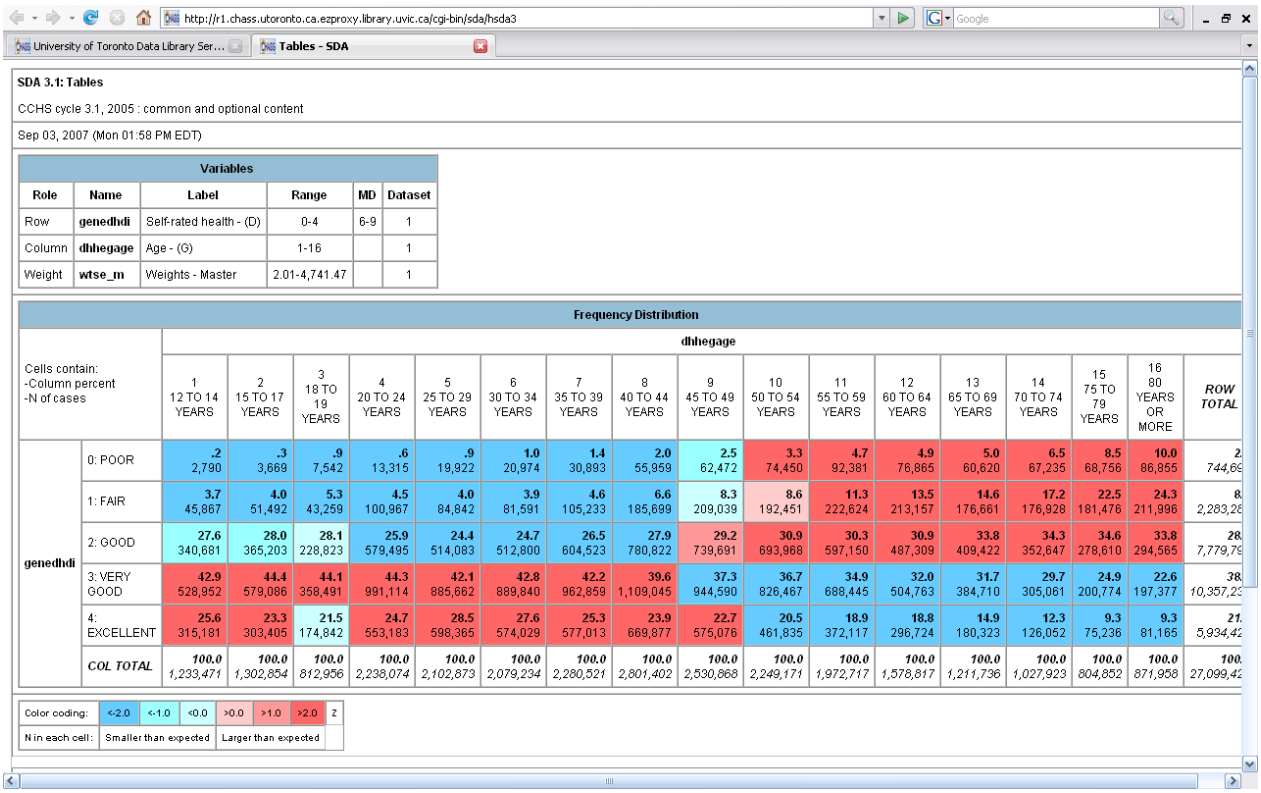
Age variable:

- Double click on the "DHH Demographics and household" variables heading.
- Double clink on the "age, sex, marital status" sub-heading.
- Click on the variable "dhhegage – Age – (G)". The variable name will appear in the **Selected** box above.
- Click on the "Col" button. The variable name will then appear in the **Column** box in the **SDA Frequencies/Cross tabulations Program.**

✓ Cross tabs are always arranged so that the data total on the independent variable's row or column, although it is conventionally the column variable. In our example the independent variable is age which we have displayed as the column variable according to convention. Consequently, ensure that in the ***TABLE OPTIONS*** box that "Column" is selected under "Percentaging".
✓ In the ***CHART OPTIONS*** box select (No Chart) next to "Type of Chart".
✓ Click "Run the Table".



The following cross tabulation table will appear:



SDA 3.1: Tables

CCHS cycle 3.1, 2005 : common and optional content

Sep 03, 2007 (Mon 01:58 PM EDT)

Variables

Role	Name	Label	Range	MD	Dataset
Row	genedhdi	Self-rated health - (D)	0-4	6-9	1
Column	dhhegage	Age - (G)	1-16		1
Weight	wtse_m	Weights - Master	2.01-4,741.47		1

Frequency Distribution

Cells contain:
-Column percent
-N of cases

dhhegage

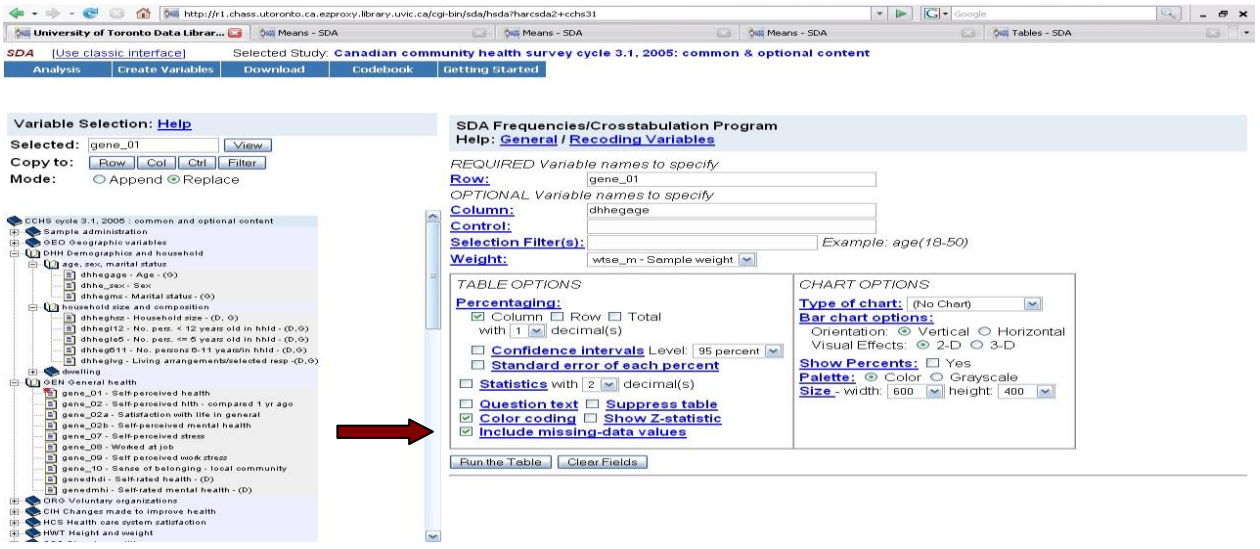
genedhdi	1 12 TO 14 YEARS	2 15 TO 17 YEARS	3 18 TO 19 YEARS	4 20 TO 24 YEARS	5 25 TO 29 YEARS	6 30 TO 34 YEARS	7 35 TO 39 YEARS	8 40 TO 44 YEARS	9 45 TO 49 YEARS	10 50 TO 54 YEARS	11 55 TO 59 YEARS	12 60 TO 64 YEARS	13 65 TO 69 YEARS	14 70 TO 74 YEARS	15 75 TO 79 YEARS	16 80 YEARS OR MORE	ROW TOTAL
0: POOR	.2 2,790	.3 3,669	.9 7,542	.6 13,315	.9 19,922	1.0 20,974	1.4 30,893	2.0 55,959	2.5 62,472	3.3 74,450	4.7 92,381	4.9 76,865	5.0 60,620	6.5 67,235	8.5 68,756	10.0 86,855	2 744,69
1: FAIR	3.7 45,867	4.0 51,492	5.3 43,259	4.5 100,967	4.0 84,842	3.9 81,591	4.6 105,233	6.6 185,699	8.3 209,039	8.6 192,451	11.3 222,624	13.5 213,157	14.6 176,661	17.2 176,928	22.5 181,476	24.3 211,996	8 2,283,26
2: GOOD	27.6 340,681	28.0 365,203	28.1 228,823	25.9 579,495	24.4 514,083	24.7 512,800	26.5 604,523	27.9 780,822	29.2 739,691	30.9 693,968	30.3 597,150	30.9 487,309	33.8 409,422	34.3 352,647	34.6 278,610	33.8 294,565	28 7,779,79
3: VERY GOOD	42.9 528,952	44.4 579,086	44.1 358,491	44.3 991,114	42.1 885,662	42.8 889,840	42.2 962,859	39.6 1,109,045	37.3 944,590	36.7 826,467	34.9 688,445	32.0 504,763	31.7 384,710	29.7 305,061	24.9 200,774	22.6 197,377	38 10,357,23
4: EXCELLENT	25.6 315,181	23.3 303,405	21.5 174,842	24.7 553,183	28.5 598,365	27.6 574,029	25.3 577,013	23.9 669,877	22.7 575,076	20.5 461,835	18.9 372,117	18.8 296,724	14.9 180,323	12.3 126,052	9.3 75,236	9.3 81,165	21 5,934,42
COL TOTAL	*100.0* *1,233,471*	*100.0* *1,302,854*	*100.0* *812,956*	*100.0* *2,238,074*	*100.0* *2,102,873*	*100.0* *2,079,234*	*100.0* *2,280,521*	*100.0* *2,801,402*	*100.0* *2,530,868*	*100.0* *2,249,171*	*100.0* *1,972,717*	*100.0* *1,578,817*	*100.0* *1,211,736*	*100.0* *1,027,923*	*100.0* *804,852*	*100.0* *871,958*	*100* *27,099,42*

Color coding: <-2.0 | <-1.0 | <0.0 | >0.0 | >1.0 | >2.0 | Z

N in each cell: Smaller than expected | Larger than expected

As we can see from the table, the data supports our hypothesis as the percentage of people with fair or poor self-rated health increases with age.

In most cases you should exclude people who did not respond to the question from your data analyses. In rare cases, they might be included if you expect that non-response is related to another variable of interest. In order to include non-responders in your cross tab select the "Include missing-data values" before running the table.

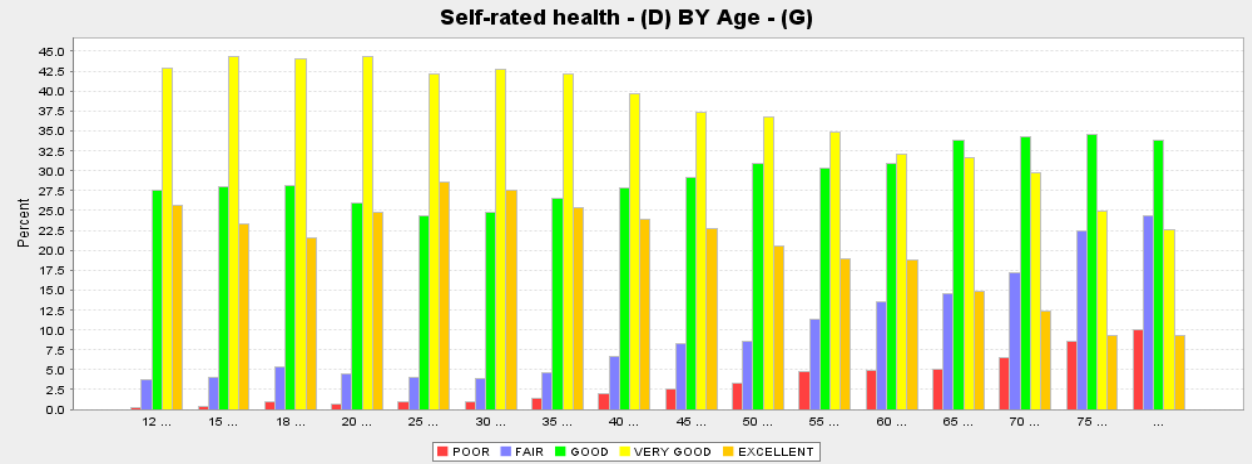


2.2 Graphing Cross tab data

We are able to display cross tab information graphically using the UT/DLS. In order to graph the cross tab created above:

- ✓ In the *CHART OPTIONS* select "Bar Chart".
- ✓ Select "Run the Table"

Scrolling past the cross tab the following chart will appear:



Notice the trend that the percentage of people rating their health as "Excellent" decreases over time, while the number who rate their health as "Poor" increases.

2.3 Calculating Chi-Square (χ^2) and other Statistics

The Pearson Chi-square is the test of statistical significance for nominal variables. It tells us whether a nominal-level association between two variables is likely to occur by chance. Chi-square is calculated from a cross tab.

To calculate the Chi-square statistic when running the above cross tab:

- ✓ In the *TABLE OPTIONS* section select "Statistics"
- ✓ Under Weight, select "No weight."
- ✓ Select "Run the table"

Below the cross tab will appear the following output:

Summary Statistics						
Eta* =	.28	Gamma =	-.23	Chisq(P) =	11,406.27	(p= 0.00)
R =	-0.26	Tau-b =	-.19	Chisq(LR) =	11,491.60	(p= 0.00)
Somers' d* =	-.17	Tau-c =	-.20	df =	60	

*Row variable treated as the dependent variable.

You will notice various other statistics are produced with this option selected include the Eta, Gamma, and

likelihood-ratio chi-square.

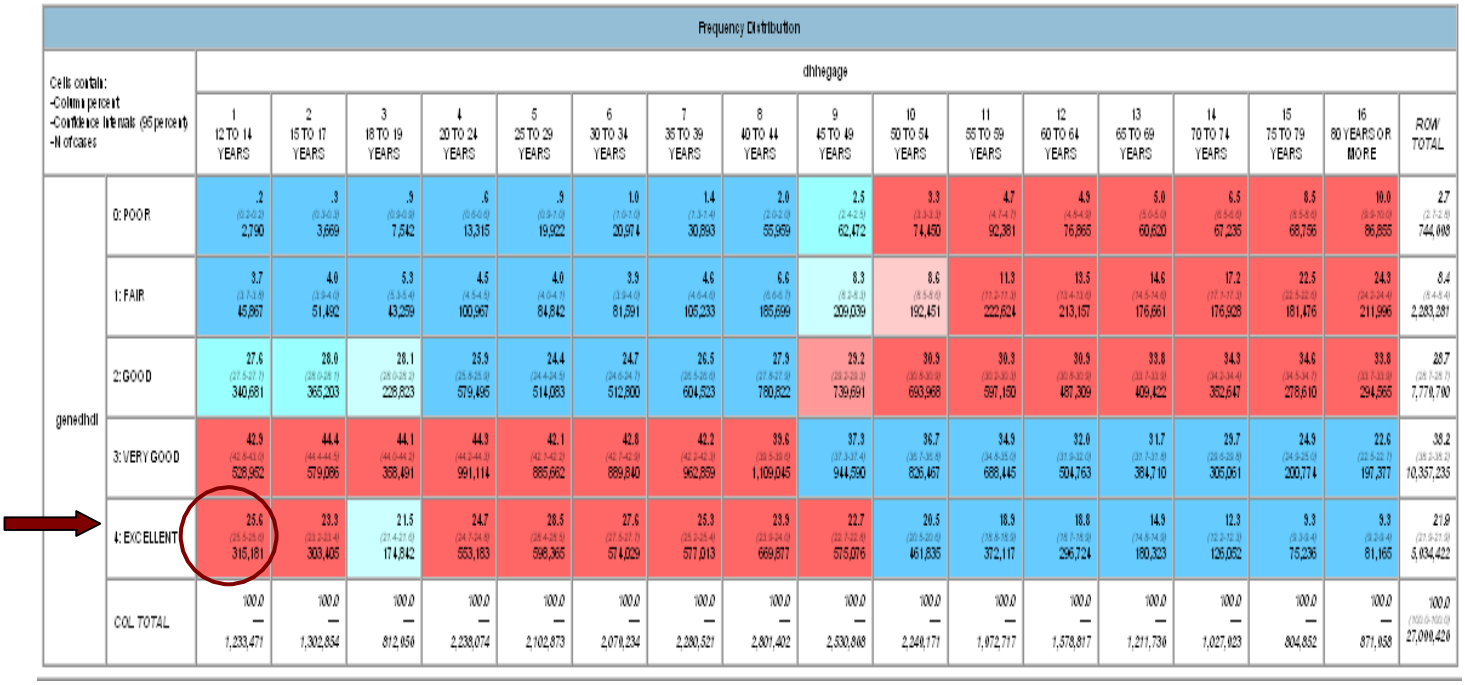
2.4 Calculating Confidence Intervals

Confidence intervals are "an indicator of the accuracy with which a population parameter can be predicted from a sample statistic" (Manheim et. al, pp 404). Confidence intervals are expressed as the range of values above and below the sample statistic the population parameter is likely to fall.

Confidence intervals can be calculated using the SDA Frequencies/Cross tabulation Program by:

- ✓ Selecting "Confidence Interval" under *TABLE OPTIONS*
- ✓ Indicating the desired Confidence Level (90%, 95% or 99%)
- ✓ Select "Run the table"

Using the example above the resulting table will be:



Frequency Distribution

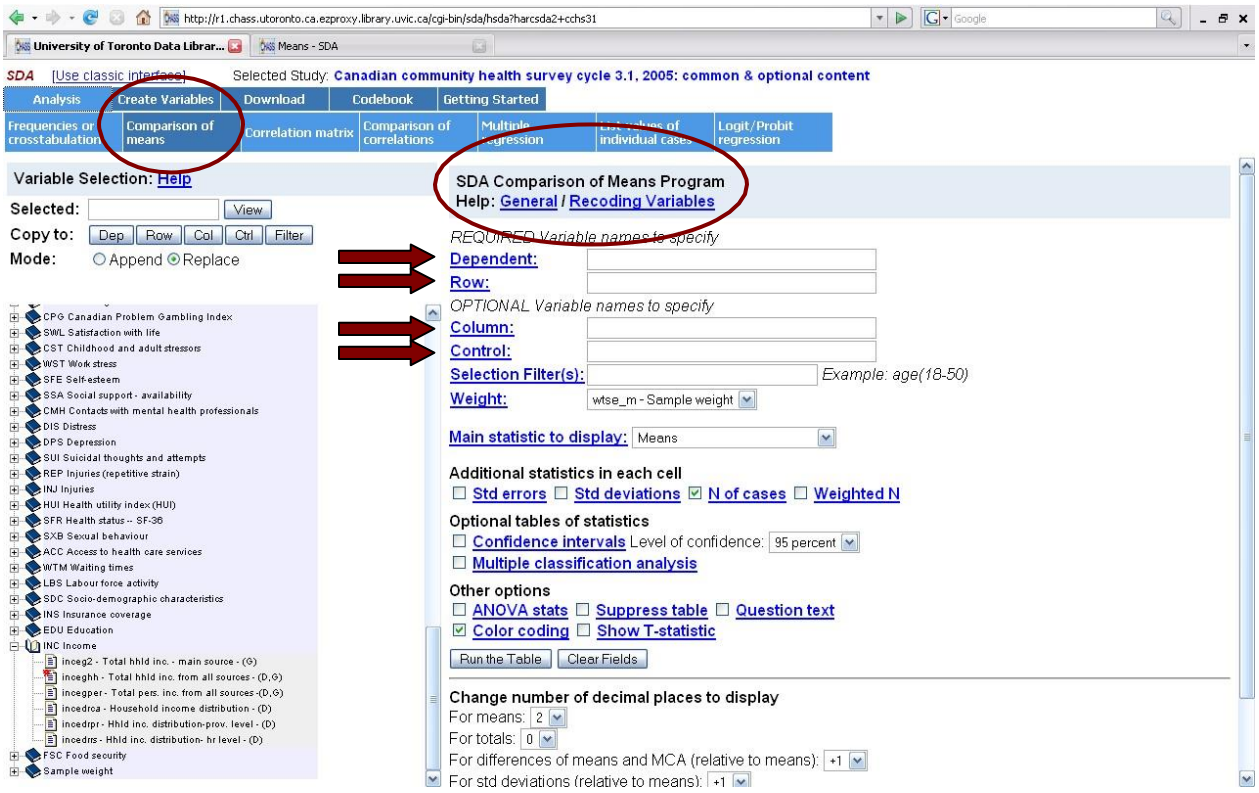
Cells contain: -Column percent -Confidence Intervals (95 percent) -N of cases		dhhegage															
		1 12 TO 14 YEARS	2 15 TO 17 YEARS	3 18 TO 19 YEARS	4 20 TO 24 YEARS	5 25 TO 29 YEARS	6 30 TO 34 YEARS	7 35 TO 39 YEARS	8 40 TO 44 YEARS	9 45 TO 49 YEARS	10 50 TO 54 YEARS	11 55 TO 59 YEARS	12 60 TO 64 YEARS	13 65 TO 69 YEARS	14 70 TO 74 YEARS	15 75 TO 79 YEARS	16 80 YEARS OR MORE	ROW TOTAL
genedhdi	0: POOR	.2 (0.2-0.2) 2,790	.3 (0.3-0.3) 3,669	.9 (0.9-0.9) 7,542	.6 (0.6-0.6) 13,315	.9 (0.9-1.0) 19,922	1.0 (1.0-1.0) 20,974	1.4 (1.3-1.4) 30,893	2.0 (2.0-2.0) 55,959	2.5 (2.4-2.5) 62,472	3.3 (3.3-3.3) 74,450	4.7 (4.7-4.7) 92,381	4.9 (4.8-4.9) 76,865	5.0 (5.0-5.0) 60,620	6.5 (6.5-6.5) 67,235	8.5 (8.5-8.5) 68,755	10.0 (9.9-10.0) 86,855	2.7 (2.7-2.8) 744,008
	1: FAIR	3.7 (3.7-3.8) 45,867	4.0 (3.9-4.0) 51,492	5.3 (5.3-5.4) 43,259	4.5 (4.5-4.5) 100,967	4.0 (4.0-4.1) 84,842	3.9 (3.9-4.0) 81,591	4.6 (4.6-4.6) 105,233	6.6 (6.6-6.7) 165,699	8.3 (8.2-8.3) 209,039	8.6 (8.5-8.6) 192,451	11.3 (11.2-11.3) 222,624	13.5 (13.4-13.6) 213,157	14.6 (14.5-14.6) 176,661	17.2 (17.1-17.3) 176,928	22.5 (22.5-22.6) 181,476	24.3 (24.2-24.4) 211,996	8.4 (8.4-8.4) 2,283,281
	2: GOOD	27.6 (27.5-27.7) 340,681	28.0 (28.0-28.1) 365,203	28.1 (28.0-28.2) 228,823	25.9 (25.8-25.9) 579,495	24.4 (24.4-24.5) 514,083	24.7 (24.6-24.7) 512,800	26.5 (26.5-26.6) 604,523	27.9 (27.8-27.9) 780,822	29.2 (29.2-29.3) 739,691	30.9 (30.8-30.9) 693,968	30.3 (30.2-30.3) 597,150	30.9 (30.8-30.9) 487,309	33.8 (33.7-33.9) 409,422	34.3 (34.3-34.4) 352,647	34.6 (34.5-34.7) 278,610	33.8 (33.7-33.9) 294,565	28.7 (28.7-28.7) 7,778,790
	3: VERY GOOD	42.9 (42.8-43.0) 528,952	44.4 (44.4-44.5) 579,086	44.1 (44.0-44.2) 368,491	44.3 (44.2-44.3) 991,114	42.1 (42.1-42.2) 885,662	42.8 (42.7-42.9) 889,840	42.2 (42.2-42.3) 962,859	39.6 (39.5-39.6) 1,109,045	37.3 (37.3-37.4) 944,590	36.7 (36.7-36.8) 826,467	34.9 (34.8-35.0) 688,445	32.0 (31.9-32.0) 504,763	31.7 (31.7-31.8) 384,710	29.7 (29.6-29.8) 305,061	24.9 (24.9-25.0) 200,774	22.6 (22.5-22.7) 197,377	38.2 (38.2-38.2) 10,357,235
	4: EXCELLENT	25.6 (25.5-25.6) 315,181	23.3 (23.2-23.4) 303,405	21.5 (21.4-21.6) 174,842	24.7 (24.7-24.8) 553,183	28.5 (28.4-28.5) 598,365	27.6 (27.5-27.7) 574,029	25.3 (25.2-25.4) 577,013	23.9 (23.9-24.0) 669,877	22.7 (22.7-22.8) 575,076	20.5 (20.5-20.6) 461,835	18.9 (18.8-18.9) 372,117	18.8 (18.7-18.9) 296,724	14.9 (14.8-14.9) 180,323	12.3 (12.2-12.3) 126,052	9.3 (9.3-9.4) 75,236	9.3 (9.2-9.4) 81,165	21.9 (21.9-21.9) 5,034,422
	COL TOTAL	100.0 — 1,233,471	100.0 — 1,302,854	100.0 — 812,956	100.0 — 2,238,074	100.0 — 2,102,873	100.0 — 2,070,234	100.0 — 2,280,521	100.0 — 2,801,402	100.0 — 2,530,868	100.0 — 2,249,171	100.0 — 1,972,717	100.0 — 1,578,817	100.0 — 1,211,736	100.0 — 1,027,923	100.0 — 804,852	100.0 — 871,058	100.0 (100.0-100.0) 27,099,426

Notice how in the cells of the table it now include the 95% confidence interval and that the confidence interval range is indicated below the column percent in the table. In this example the confidence interval indicates that 95% percent of such intervals calculated, the percentage of the population that falls within the specific category (e.g. 12-14 years olds with Excellent self-rated health) will occur within the indicated range (e.g. 25.5% - 25.6%).

3. Comparisons of Means

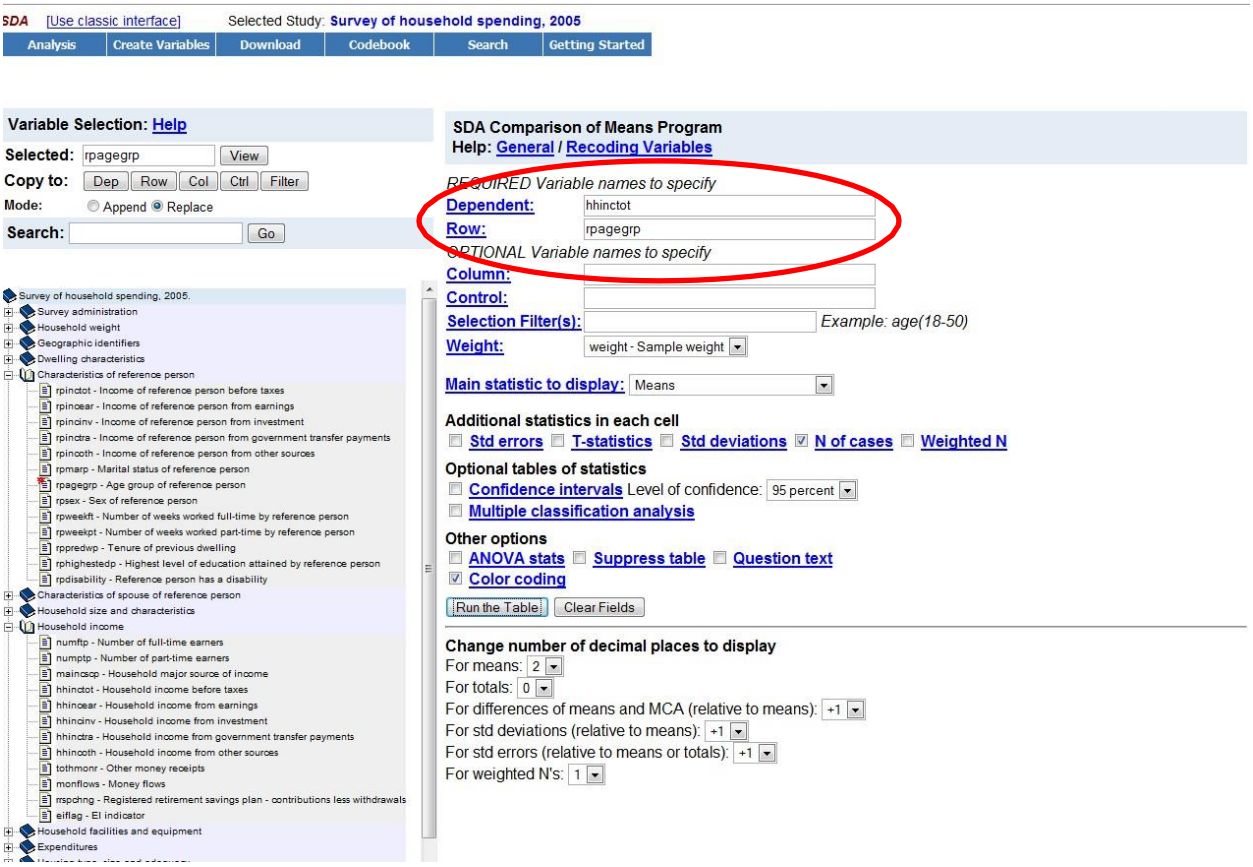
Cross tabulations are suitable for the examination of a relationship between two nominal/ordinal level variables. But what if your variable of interest is an interval or ratio level variable? The **SDA Comparison of Means** analysis program is able to calculate means of variables and can also do so separately within categories of a selected independent variable and, optionally, a

selected column variable. If a control variable is also specified, a separate table will be produced for each category of the control variable. A more in-depth explanation of each option can be obtained by selecting the corresponding word highlighted on the SDA Comparison of Means Program.



In order to demonstrate how the program is used we will examine the impact of age on income using the Survey of household spending (SHS) 2005. In the SHS, income is a ratio level coded variable.

- ✓ Select "Data" next to 2005
- ✓ On the toolbar at the top of the page highlight "Analysis" and select the "Comparison of means" program.
- ✓ Using the Variable Selection program double click on the "Household income" variables heading and copy the variable "hhinctot" into the Dependent (Dep) variable box.
- ✓ Using the Variable Selection program double click on the "Characteristics of reference person" variables heading and copy the variable "rpagegrp" into the Row (Row) variable box. Age is interval coded in the SHS.
- ✓ **Optional:** If you also want to calculate standard errors, confidence intervals, or other statistics, select the appropriate options.
- ✓ Click "Run the Table"



The following table will be created:

Variables

Role	Name	Label	Range	MD	Dataset
Dependent	**hhinctot**	Household income before taxes	-17,000.00-1,900,000.00		1
Row	**rpagegrp**	Age group of reference person	1-14		1
Weight	**weight**	Weight at household level	10-8479		1

Main Statistics

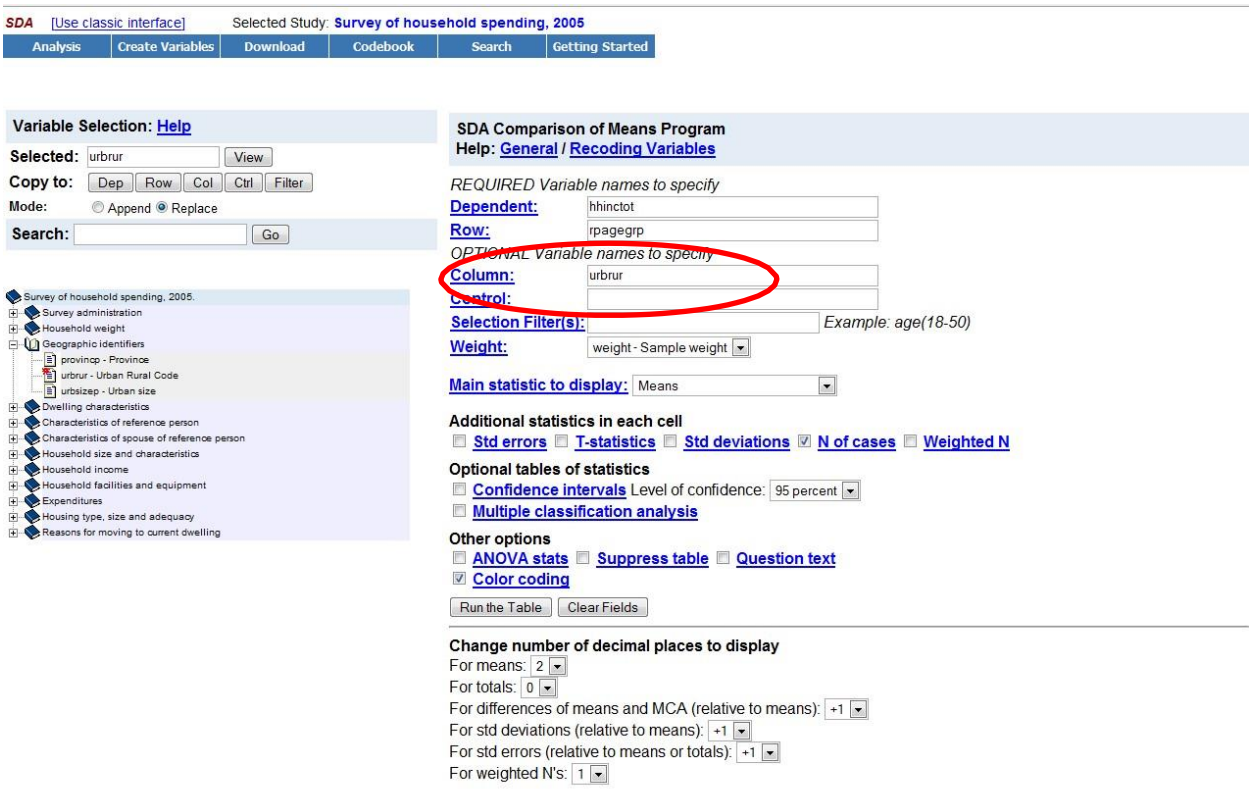
Cells contain: -Means -N of cases		
rpagegrp	1: Less than 25 years	**35,825.20** 701
	2: 25-29 years	**53,631.09** 1,026

	3: 30-34 years	**68,005.75** 1,213
	4: 35-39 years	**79,240.06** 1,403
	5: 40-44 years	**79,161.75** 1,774
	6: 45-49 years	**84,032.90** 1,672
	7: 50-54 years	**83,837.96** 1,649
	8: 55-59 years	**74,227.10** 1,476
	9: 60-64 years	**67,931.61** 1,098
	10: 65-69 years	**48,632.16** 888
	11: 70-74 years	**39,621.42** 823
	12: 75-79 years	**34,502.24** 673
	13: 80-84 years	**32,268.54** 492
	14: 85 years and over	**29,329.70** 334
	COL TOTAL	***66,654.87*** *15,222*

From the above results we see that average income rises with age, peaking in the 45-49 age group and then declines.

We can also compare means for additional variables in order to examine relationships between multiple variables.

- ✓ Using the Variable Selection program double click on the "Geographic identifiers" variables heading and copy the variable "urbrur", which is an indicator for whether the respondent lives in a rural or urban area, into the Column (Col) variable box.
- ✓ Click "Run the Table"



The following table will be created:

Variables

Role	Name	Label	Range	MD	Dataset
Dependent	**hhinctot**	Household income before taxes	-17,000.00-1,900,000.00		1
Row	**rpagegrp**	Age group of reference person	1-14		1
Column	**urbrur**	Urban Rural Code	1-2	0	1
Weight	**weight**	Weight at household level	10-8479		1

Main Statistics

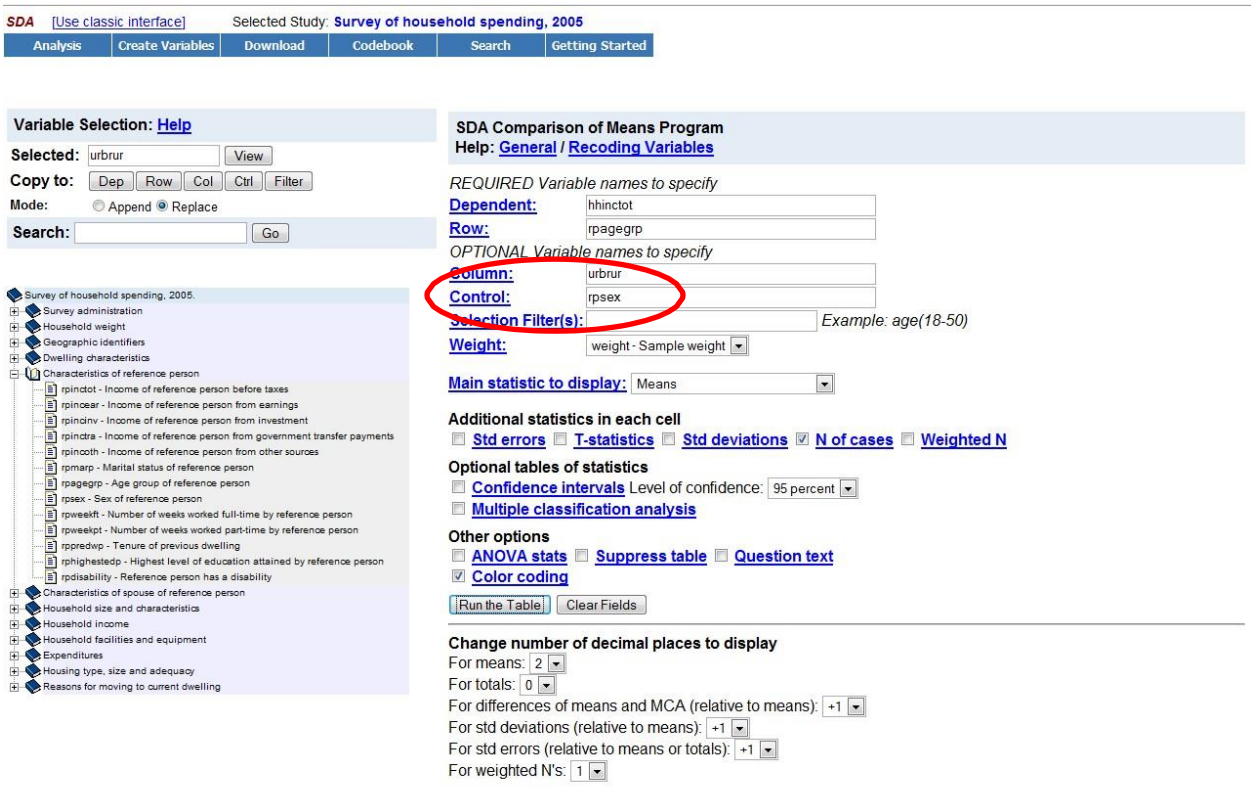
Cells contain: -Means -N of cases		**urbrur**		
		1 Urban	2 Rural	***ROW TOTAL***
rpagegrp	1: Less than 25 years	**35,782.42** 622	**36,617.09** 60	***35,835.03*** *682*
	2: 25-29 years	**53,643.32**	**53,951.47**	***53,662.65***

		869	116	*985*
	3: 30-34 years	**67,747.73** 1,033	**71,785.65** 137	***68,028.70*** *1,170*
	4: 35-39 years	**80,541.69** 1,141	**65,394.00** 206	***78,861.30*** *1,347*
	5: 40-44 years	**81,108.64** 1,416	**64,769.08** 277	***79,206.08*** *1,693*
	6: 45-49 years	**86,307.88** 1,313	**65,659.35** 297	***83,754.34*** *1,610*
	7: 50-54 years	**86,000.36** 1,274	**61,510.32** 303	***82,676.86*** *1,577*
	8: 55-59 years	**77,299.62** 1,149	**56,767.07** 264	***74,340.07*** *1,413*
	9: 60-64 years	**67,504.16** 824	**48,767.14** 216	***64,613.18*** *1,040*
	10: 65-69 years	**48,115.92** 664	**37,184.21** 175	***46,571.04*** *839*
	11: 70-74 years	**40,093.65** 634	**37,007.51** 147	***39,655.21*** *781*
	12: 75-79 years	**35,284.05** 502	**29,342.13** 145	***34,515.34*** *647*
	13: 80-84 years	**32,385.25** 371	**31,754.88** 99	***32,282.55*** *470*
	14: 85 years and over	**30,359.70** 261	**21,035.81** 61	***29,378.83*** *322*
	COL TOTAL	***67,728.52*** *12,073*	***54,400.48*** *2,503*	***66,152.07*** *14,576*

From the comparison of means we can see that average income is higher for rural respondents who are younger than 35 but past 35 years of age, average income is higher for urban respondents.

We can also control for other variables. When a control variable is used tables are printed for every value of the control variable (e.g. male and female). To control for gender:

- ✓ Using the Variable Selection program double click on the "Characteristics of reference person" variables heading and copy the variable "rpsex" into the Control variable box.
- ✓ Click "Run the Table"



Three tables will be created, one for male respondents, one for female respondents, and one for all valid cases.

Variables

Role	Name	Label	Range	MD	Dataset
Dependent	**hhinctot**	Household income before taxes	-17,000.00-1,900,000.00		1
Row	**rpagegrp**	Age group of reference person	1-14		1
Column	**urbrur**	Urban Rural Code	1-2	0	1
Control	**rpsex**	Sex of reference person	1-2		1
Weight	**weight**	Weight at household level	10-8479		1

Statistics for rpsex = 1(Male)

Cells contain: -Means -N of cases	**urbrur**		
	1 Urban	2 Rural	***ROW TOTAL***

rpagegrp	1: Less than 25 years	**35,873.68** 260	**33,445.03** 21	***35,719.62*** *281*

2: 25-29 years	**51,898.02** 389	**57,417.53** 48	***52,206.37*** *437*
3: 30-34 years	**67,967.85** 465	**83,269.13** 55	***69,118.79*** *520*
4: 35-39 years	**75,938.69** 530	**69,919.61** 76	***75,447.13*** *606*
5: 40-44 years	**83,559.56** 664	**67,512.70** 134	***81,740.47*** *798*
6: 45-49 years	**86,560.35** 626	**64,483.45** 151	***83,647.90*** *777*
7: 50-54 years	**89,670.86** 662	**60,414.92** 161	***85,427.12*** *823*
8: 55-59 years	**87,874.29** 582	**65,643.35** 139	***84,540.24*** *721*
9: 60-64 years	**75,416.40** 449	**53,025.62** 110	***71,995.36*** *559*
10: 65-69 years	**57,012.50** 340	**42,088.20** 85	***54,925.66*** *425*
11: 70-74 years	**46,509.84** 291	**39,895.20** 81	***45,478.25*** *372*
12: 75-79 years	**42,727.26** 219	**32,758.89** 60	***41,390.45*** *279*
13: 80-84 years	**42,365.49** 137	**33,189.00** 45	***40,716.83*** *182*
14: 85 years and over	**45,109.38** 83	**29,786.12** 26	***43,095.91*** *109*
COL TOTAL	***71,937.14*** *5,697*	***57,888.52*** *1,192*	***70,254.45*** *6,889*

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	**T**
Mean in each cell:	Smaller than average			Larger than average			

Statistics for rpsex = 2(Female)

Cells contain: -Means -N of cases	**urbrur**		
	1 Urban	2 Rural	***ROW TOTAL***

rpagegrp	1: Less than 25 years	**35,708.99**	**39,201.07**	***35,927.96***

		362	39	*401*
	2: 25-29 years	**55,329.48** 480	**51,289.88** 68	***55,049.59*** *548*
	3: 30-34 years	**67,547.73** 568	**59,456.98** 82	***67,026.66*** *650*
	4: 35-39 years	**85,938.07** 611	**62,565.35** 130	***82,596.79*** *741*
	5: 40-44 years	**78,817.16** 752	**62,347.91** 143	***76,852.42*** *895*
	6: 45-49 years	**86,058.48** 687	**67,013.12** 146	***83,861.49*** *833*
	7: 50-54 years	**81,603.62** 612	**63,079.61** 142	***79,302.22*** *754*
	8: 55-59 years	**66,244.23** 567	**46,536.03** 125	***63,525.42*** *692*
	9: 60-64 years	**58,157.98** 375	**43,861.53** 106	***55,926.82*** *481*
	10: 65-69 years	**38,427.38** 324	**31,978.88** 90	***37,505.63*** *414*
	11: 70-74 years	**34,083.48** 343	**33,621.84** 66	***34,024.10*** *409*
	12: 75-79 years	**28,763.87** 283	**26,102.63** 85	***28,430.73*** *368*
	13: 80-84 years	**25,957.97** 234	**30,624.26** 54	***26,666.10*** *288*
	14: 85 years and over	**24,309.77** 178	**15,802.09** 35	***23,510.13*** *213*
	COL TOTAL	***63,579.84*** *6,376*	***50,862.66*** *1,311*	***62,094.44*** *7,687*

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	**T**
Mean in each cell:	Smaller than average			Larger than average			

Statistics for all valid cases

Cells contain: -Means -N of cases	**urbrur**		
	1 Urban	2 Rural	***ROW TOTAL***

rpagegrp	1: Less than 25 years	**35,782.42** 622	**36,617.09** 60	***35,835.03*** *682*
	2: 25-29 years	**53,643.32** 869	**53,951.47** 116	***53,662.65*** *985*
	3: 30-34 years	**67,747.73** 1,033	**71,785.65** 137	***68,028.70*** *1,170*
	4: 35-39 years	**80,541.69** 1,141	**65,394.00** 206	***78,861.30*** *1,347*
	5: 40-44 years	**81,108.64** 1,416	**64,769.08** 277	***79,206.08*** *1,693*
	6: 45-49 years	**86,307.88** 1,313	**65,659.35** 297	***83,754.34*** *1,610*
	7: 50-54 years	**86,000.36** 1,274	**61,510.32** 303	***82,676.86*** *1,577*
	8: 55-59 years	**77,299.62** 1,149	**56,767.07** 264	***74,340.07*** *1,413*
	9: 60-64 years	**67,504.16** 824	**48,767.14** 216	***64,613.18*** *1,040*
	10: 65-69 years	**48,115.92** 664	**37,184.21** 175	***46,571.04*** *839*
	11: 70-74 years	**40,093.65** 634	**37,007.51** 147	***39,655.21*** *781*
	12: 75-79 years	**35,284.05** 502	**29,342.13** 145	***34,515.34*** *647*
	13: 80-84 years	**32,385.25** 371	**31,754.88** 99	***32,282.55*** *470*
	14: 85 years and over	**30,359.70** 261	**21,035.81** 61	***29,378.83*** *322*
	COL TOTAL	***67,728.52*** *12,073*	***54,400.48*** *2,503*	***66,152.07*** *14,576*

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	**T**
Mean in each cell:	Smaller than average			Larger than average			

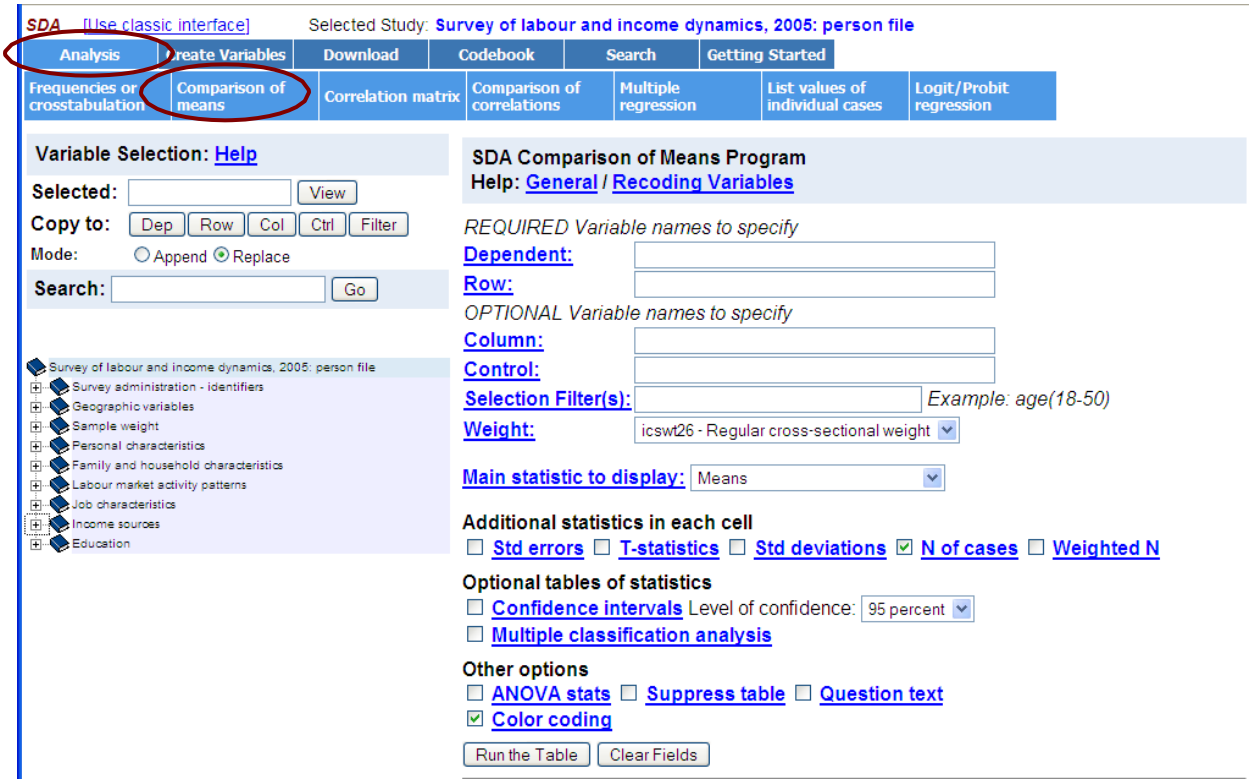
If you only want to analyze a sub-group of a variable (e.g. female respondents) apply Selection Filter criteria in the same way as in the SDA Frequencies/Crosstabulation Program.

3.1 T-Test: Test of Overall Mean

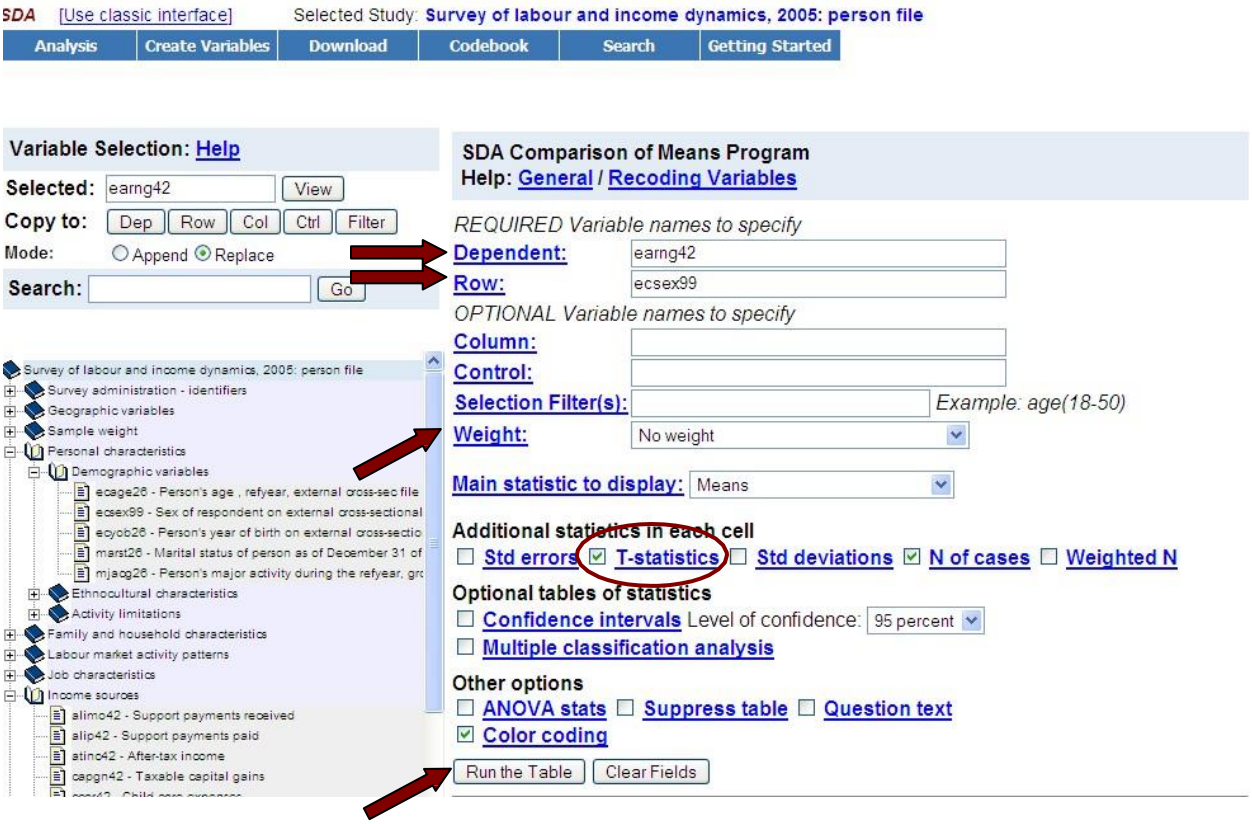
The *t*-statistic shows whether the mean in a cell is larger or smaller than the overall mean. It also takes into account the total number of cases in the cell. If there are only a few cases in a cell, the deviation from the overall mean is not as significant as if there are many cases in that cell.

The example of the t-test will examine the relationship between gender and earnings. For this example we will examine data from the Survey of Labour and Income Dynamics (SLID) from 2005, using the person file.

- ✓ Select "Data" next to person file for 2005
- ✓ On the toolbar at the top of the page highlight "Analysis" and select the "Comparison of means" program.



- ✓ Using the Variable Selection program double click on "Income sources" and select the "Earnings" (earng42) variable. Copy this variable into the Dependent (Dep) variable box.
- ✓ Using the Variable Selection program double click on the "Personal characteristics" variables heading, then expand the "Demographic variables" subheading and copy the variable "Sex of respondent on external cross-sectional files" (ecsex99) into the Row (Row) variable box.
- ✓ Under Weight, select "No weight."
- ✓ Under 'Additional statistics in each cell' select "T-statistic."



✓ Click "Run the Table"

SDA 3.2: Means

Survey of labour and income dynamics, 2005: person file

Oct 04, 2008 (Sat 04:38 PM EDT)

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Dependent	**earng42**	Earnings	-82500-1700000		1
Row	**ecsex99**	Sex of respondent on external cross-sectional files	1-2	6-9	1

Main Statistics		
Cells contain: -Means -N of cases -T-statistic		
ecsex99	1: Male	**31,504.98** 25,581 26.8
	2: Female	**17,431.78** 27,893 -46.6
	COL TOTAL	***24,164.15*** *53,474* ---

Mean earnings (Male)

N of cases (Male)

T-statistic

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	T
Mean in each cell:	Smaller than average			Larger than average			

Explanation of colour coding

Allocation of cases	
Valid cases	53,474
Total cases	*53,474*

Datasets	
1	/dli2/slid/slid05pr
2	/newvars/slid/slid05pr

CSM, UC Berkeley

The results provide a range of information, as seen in the diagram above. The t-statistic is provided as a number and affects the colour coding of the table of means. Below the table is an explanation of the colour coding.

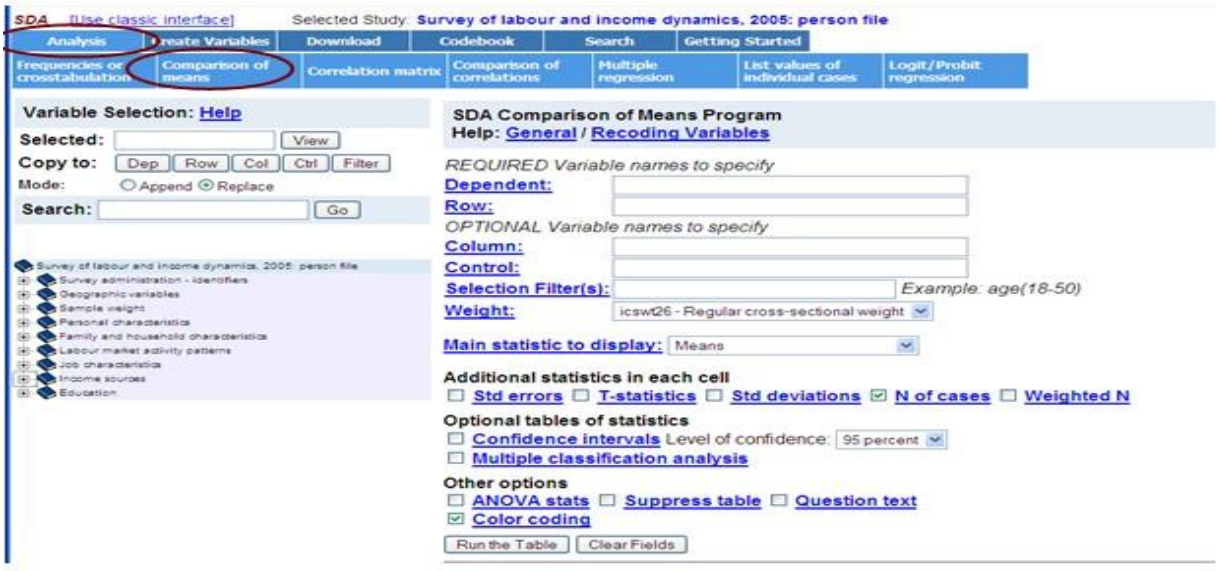
From the comparison of means t-test we can see that the mean earnings is higher than the overall mean and is lower than the overall mean for female respondents. Use the t-test table to find the critical value at the 95% confidence level (0.05) for a sample of 53,474. The critical value is 1.96. Therefore, the results of this analysis are statistically significant at a 95% confidence level.

3.2 T-Test: Test of Mean Between Subgroups

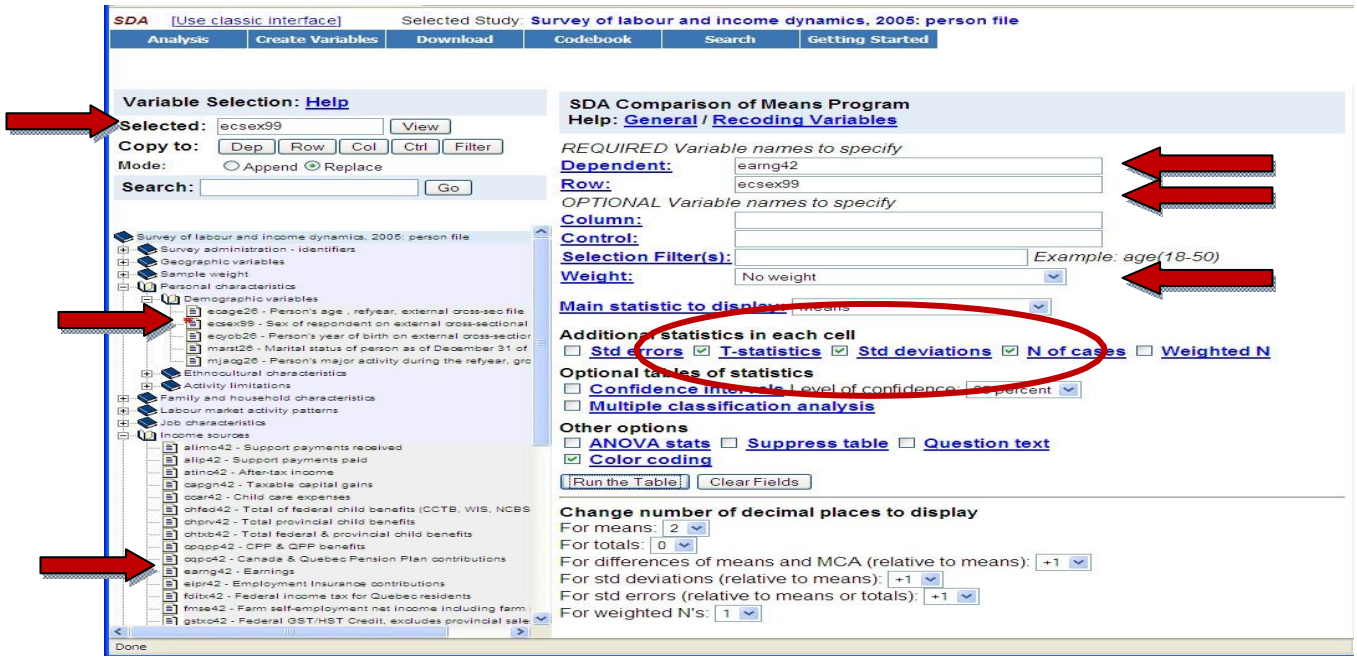
This test assesses if the mean of each subgroup is different from each other. This test requires one to perform the calculations oneself and use a t-table.
The example of the t-test will examine the relationship between gender and earnings. For this example we will examine data from the Survey of Labour and Income Dynamics (SLID) from 2005, using the person file where measure of income earnings is ratio level.

✓ Select "Data" next to person file for 2005

✓ On the toolbar at the top of the page highlight "Analysis" and select the "Comparison of means" program.



- ✓ Using the Variable Selection program double click on “Income sources” and select the “Earnings” (earng42) variable. Copy this variable into the Dependent (Dep) variable box.
- ✓ Using the Variable Selection program double click on the “Personal characteristics” variables heading, then expand the “Demographic variables” subheading and copy the variable “Sex of respondent on external cross-sectional files” (ecsex99) into the Row (Row) variable box.
- ✓ Under Weight, select “No weight.”
- ✓ Select t-statistic, Std deviation (standard deviation), and N of cases.



- ✓ Click “Run the Table”

SDA 3.2: Means

Survey of labour and income dynamics, 2005: person file

Nov 07, 2008 (Fri 08:55 PM EST)

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Dependent	**earng42**	Earnings	-82500-1700000		1
Row	**ecsex99**	Sex of respondent on external cross-sectional files	1-2	6-9	1

Main Statistics		
Cells contain: -Means -Std Devs -N of cases -T-statistic		
ecsex99	1: Male	**31,504.98** 43,772.874 25,581 26.8
	2: Female	**17,431.78** 24,149.639 27,893 -46.6
	COL TOTAL	***24,164.15*** *35,640.136* *53,474* ---

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	T
Mean in each cell:	Smaller than average			Larger than average			

The result shows that women earn about half of what men do. To check if this difference is significant a further calculation of the t-test statistic is required.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}} \quad (\text{where d.f.} = n_1 + n_2 - 2)$$

where:

s_i^2 = variance of X for category i
n_i = sample size category i
X_i = sample mean of variable X for category i

The calculation is as follows:

$$t = \frac{31504.98 - 24164.15}{\sqrt{\frac{43772.874^2}{25581-1} + \frac{35640.136^2}{53474-1}}}$$

$$= \frac{7340.83}{\sqrt{74904.78882 +}}$$

$$= \frac{7340.83}{314.1006108}$$

$$= 23.38$$

$$\text{d.f.} = \infty$$

The answer of 23.38 is then compared to the critical value. Use the t-test table to find the critical value at the 95% confidence level (0.05) for a sample of 53,474. The critical value is 1.96. We therefore easily reject the null in favour of the alternative as 23.38>1.96.

4. Correlations

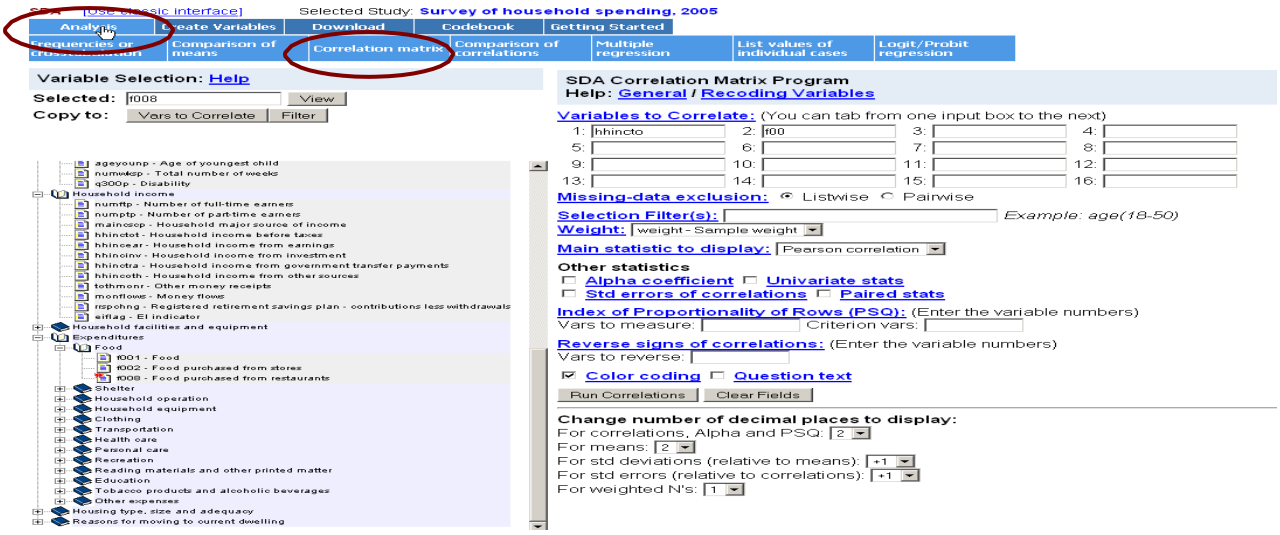
Correlation measures the relative strength of the linear relationship between two ratio level variables. The correlation coefficient, or Pearson correlation, is a measure of association between two variables. The Pearson correlation coefficient is;

- Unit free
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

The SDA Correlation Matrix Program calculates the correlation coefficient between all pairs of two or more variables.

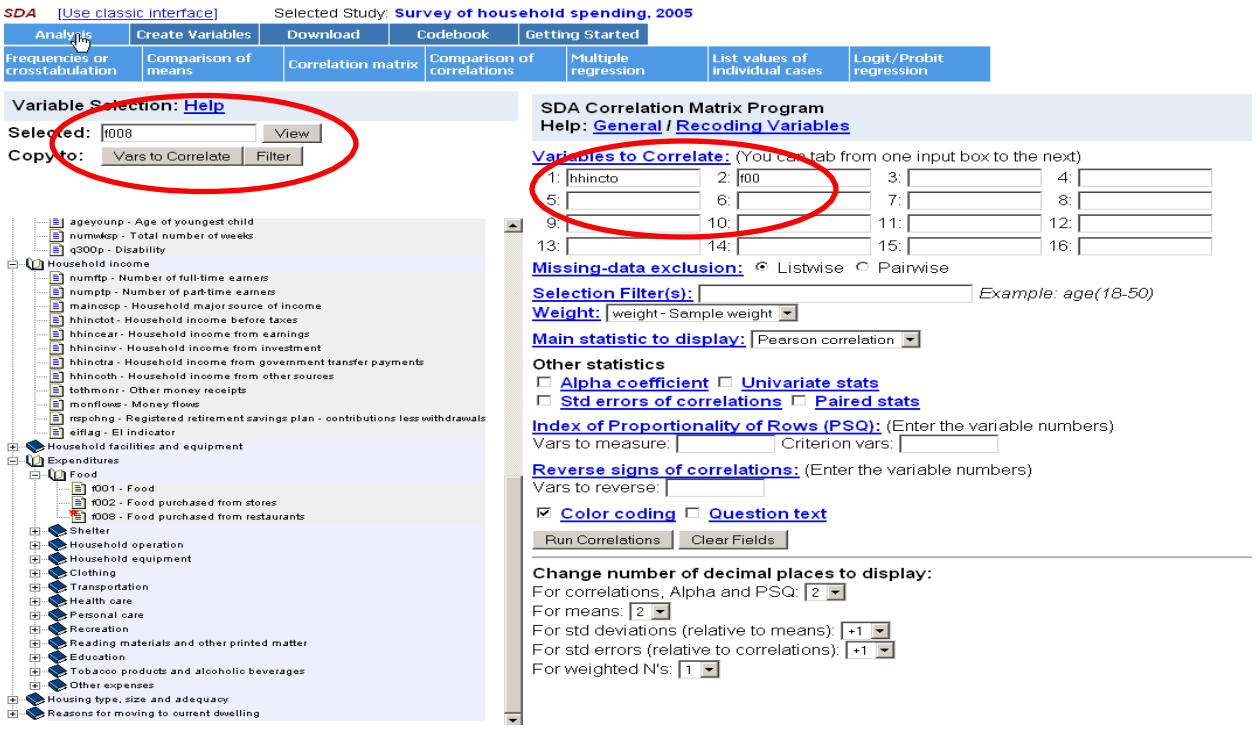
Using the Survey of Household Spending (SHS) data for 2005, we will calculate the correlation between the variables household income before tax and restaurant expenditures.

- ✓ Open the Data section of the SHS, 2005
- ✓ Select the "Analysis" tab and open the Correlation matrix program



Using the Variable Selection tool select the following variables and copy them to "Variables to Correlate":

- ✓ Double click on the "Household income" variables heading and left click on the the variable "hhinctot". Left click on the "Vars to Correlate" button.
- ✓ Double click on the "Expenditures" variables heading, then double click on the "Food" variables sub-heading, and left click on the variable "f008". Left click on the "Vars to Correlate" button.



- ✓ Beside "Main statistic to display" select "Pearson correlation"
- ✓ Select "Run Correlations"

The following table will be created:

Variables

Role	Name	Label	Range	MD	Dataset
Correlate	**hhinctot**	Household income before taxes	-17,000.00-1,900,000.00		1
Correlate	**f008**	Food purchased from restaurants	.00-29,200.00		1
Weight	**weight**	Weight at household level	10-8479		1

Correlation Matrix

	hhinctot	f008
hhinctot	1.00	.42
f008	.42	1.00

Note from the resulting matrix that there is a moderate positive correlation (.42) between before tax household income and restaurant expenditures.

The SDA Correlation Matix program allows you to explore the correlations between up to 16 variables at a time.

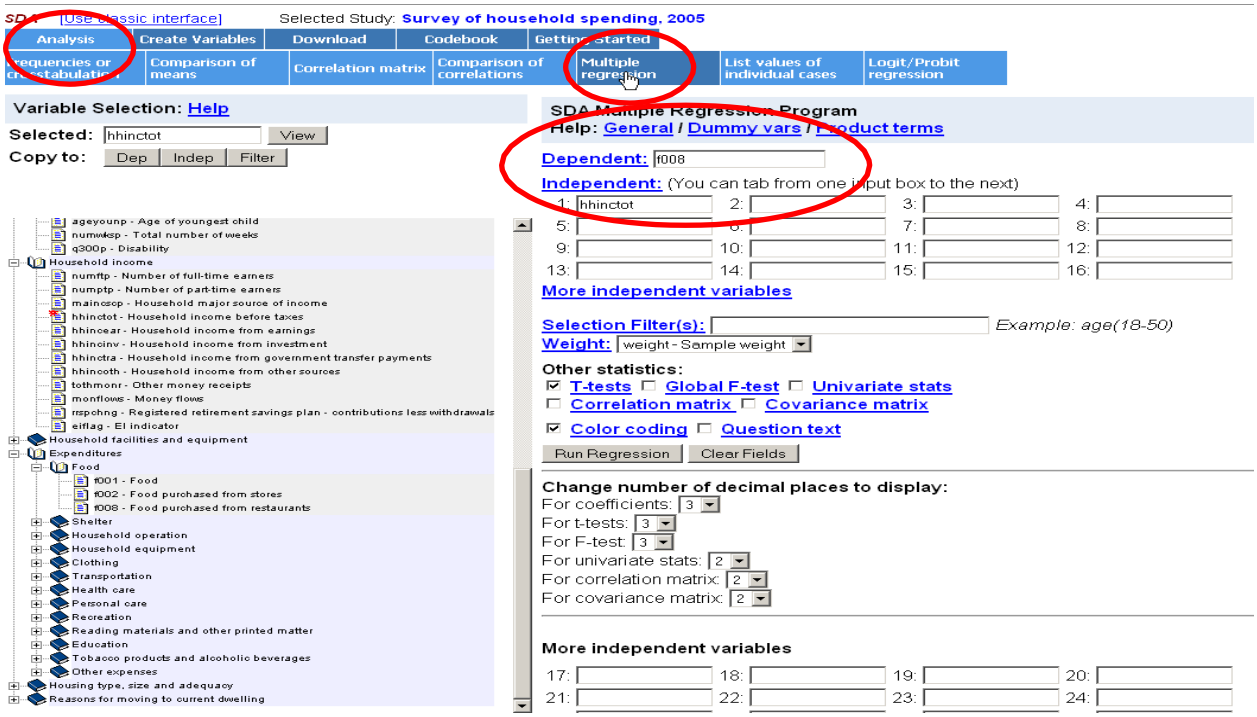
5. Simple Regressions

Regression analysis is a tool used to predict the value of a dependent variable (the variable we wish to explain or Y) based on the value of at least one independent variable (the variable used to explain independent variable or X), and explain the impact of changes in an independent variable on the dependent variable.

The SDA Multiple Regression program calculates the regression coefficients for one or more independent or predictor variables, using ordinary least squares. In this course we will only be performing simple regressions (regressing Y on X) so using the SDA multiple regression program we will only be regressing the dependent variable on a single independent variable.

We will perform a simple regression using the variables household income before tax (hhinctot) and restaurant expenditures (f008) from the SHS, 2005.

- ✓ Open the Data section of the SHS, 2005
- ✓ Select the "Analysis" tab and open the Multiple regression program
- ✓ Since we are trying to explain respondent's food expenditures using household before tax income, f008 is the dependent variable and hhinctot is the independent variable. Using the variable selection tool copy the variables into the Dependent and Independent variable positions respectively.



✓ Select "Run Regression"

The following output will be produced:

Variables

Role	Name	Label	Range	MD	Dataset
Dependent	**f008**	Food purchased from restaurants	.00-29,200.00		1
Independent	**hhinctot**	Household income before taxes	-17,000.00-1,900,000.00		1
Weight	**weight**	Weight at household level	10-8479		1

Regression Coefficients					Test That Each Coefficient = 0	
	B	**SE(B)**	**Beta**	**SE(Beta)**	**T-statistic**	**Probability**
hhinctot	.013	.000	.423	.007	57.518	.000
Constant	749.757	21.292			35.213	.000

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	T
Effect of each variable:	Negative			Positive			

Multiple R = .423 R-Squared = .179 Std Error of Estimate = 51,573.605

Two versions of the regression coefficient are given for each variable:

1. The **unstandardized** regression coefficient -- labeled ***B***
2. The **standardized** regression coefficient -- labeled ***Beta***

For each version of the coefficient there is also a **standard error** -- labeled either as ***SE(B)*** or as ***SE(Beta).*** The calculation of these standard errors assumes that the dataset is a simple random sample drawn from the target population. If the sample is more complex, the displayed standard errors may be too small.

6. Conclusion

This tutorial covers the basic material for bivariate analysis using the UT/DLS service. While multiple regression and logit/probit regression was not explicitly covered, the skills learned through this user guide should allow individuals to explore these tools on their own. In the next section, we will cover basic data manipulation skills include creating new variables, recoding existing variables, and downloading data for use in an alternate program (e.g. Excel). To ensure that you understand the basic material presented in this section, please work through the following exercises.

7. Exercises

a) Using the General Social Survey (GSS) on Victimization (Cycle 18, 2004), create a cross tabulation table and graph to examine the relationship between gender and physical or sexual violence by an ex-spouse or partner.
b) Using the Survey of Household Spending (SHS), 2005, analyze the impact of sex on reference persons food, shelter, clothing, and personal care expenditures using a comparison of means.
c) Using the Survey of Household Spending (SHS), 2005, calculate the correlation matrix for food, shelter, clothing, and personal care expenditures and interpret identified correlations.

8. Answers

a)

- ✓ Open the Data section of GSS Cycle 18.
- ✓ Double click the "Demographic characteristics" variables heading and select the variable "sex – Sex of respondent".
- ✓ Beside "Copy to" click the "Col" button.
- ✓ Double click the variables heading "Section 4: abuse by ex-spouse/partner".
- ✓ Double click the sub-heading "Module: physical and sexual violence by ex-spouse/partner".
- ✓ Select the variable "exviol – Physical or sexual violence by ex-spouse/partner".
- ✓ Beside "Copy to" click the "Row" button.
- ✓ In the *CHART OPTIONS* select "Bar Graph"
- ✓ Select "Run the Table"
- ✓ The completed cross tab should appear as follows.

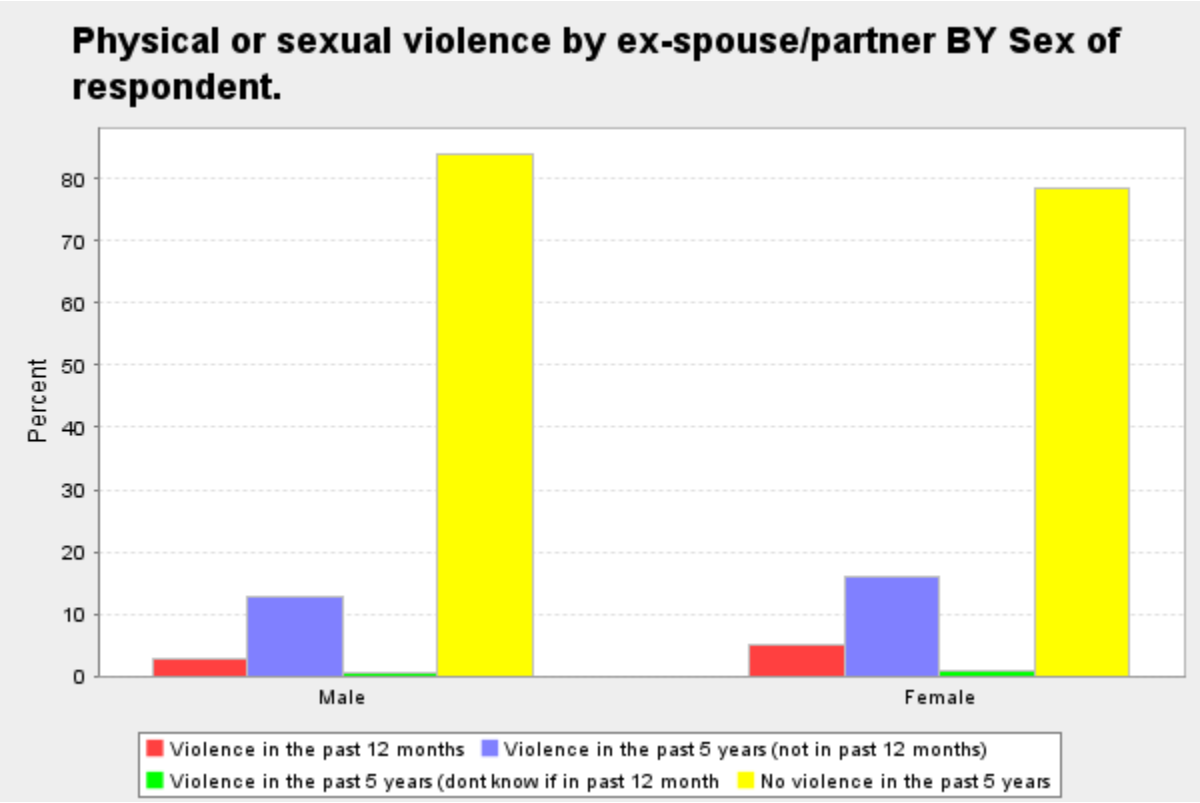
Variables

Role	Name	Label	Range	MD	Dataset
Row	**exviol**	Physical or sexual violence by ex-spouse/partner	1-4	7-*	1
Column	**sex**	Sex of respondent.	1-2		1
Weight	**wght_per**	Person weight.	36.5572-6,093.4390		1

Frequency Distribution

Cells contain: -Column percent -N of cases		**sex**		
		1 Male	2 Female	***ROW TOTAL***
exviol	1: Violence in the past 12 months	**2.9** 44,830	**5.0** 95,246	***4.1*** *140,076*
	2: Violence in the past 5 years (not in past 12 months)	**12.7** 195,470	**15.9** 302,516	***14.5*** *497,986*
	3: Violence in the past 5 years (don't know if in past 12 month	**.5** 7,104	**.7** 14,270	***.6*** *21,374*
	4: No violence in the past 5 years	**83.9** 1,289,537	**78.4** 1,494,646	***80.9*** *2,784,184*
	COL TOTAL	***100.0*** *1,536,942*	***100.0*** *1,906,678*	***100.0*** *3,443,620*

And the graph as follows:



b)

- ✓ Open the Data section of SHS 2005
- ✓ Go to "Analysis" on the menu bar and select the "Comparison of means" program
- ✓ In the Row variable box copy "rpsex – Sex of reference person"
- ✓ In the Dependent variable box APPEND the variables: "f001 – Food"; "g001 – Shelter"; "j001 – Clothing"; and "l201 – Personal care".
- ✓ Select "Run the Table"
- ✓ The completed cross tab should appear as follows.

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Dependent	**f001**	Food	1.00-64,000.00		1
Row	**rpsex**	Sex of reference person	1-2		1
Weight	**weight**	Weight at household level	10-8479		1

Main Statistics		
Cells contain: -Means -N of cases		
rpsex	1: Male	**7,288.51**

		7,151
	2: Female	**6,666.11** 8,071
	COL TOTAL	***6,975.40*** *15,222*

Variables

Role	Name	Label	Range	MD	Dataset
Dependent	**g001**	Shelter	.00-166,360.00		1
Row	**rpsex**	Sex of reference person	1-2		1
Weight	**weight**	Weight at household level	10-8479		1

Main Statistics

Cells contain: -Means -N of cases		
rpsex	1: Male	**12,558.91** 7,151
	2: Female	**12,177.80** 8,071
	COL TOTAL	***12,367.19*** *15,222*

Variables

Role	Name	Label	Range	MD	Dataset
Dependent	**j001**	Clothing	.00-69,300.00		1
Row	**rpsex**	Sex of reference person	1-2		1
Weight	**weight**	Weight at household level	10-8479		1

Main Statistics

Cells contain: -Means -N of cases		
rpsex	1: Male	**2,543.74** 7,151

	2: Female	**2,524.93** 8,071
	COL TOTAL	***2,534.28*** *15,222*

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Dependent	**l201**	Personal care	.00-11,700.00		1
Row	**rpsex**	Sex of reference person	1-2		1
Weight	**weight**	Weight at household level	10-8479		1

Main Statistics

Cells contain: -Means -N of cases		
rpsex	1: Male	**1,034.21** 7,151
	2: Female	**1,105.24** 8,071
	COL TOTAL	***1,069.94*** *15,222*

c)

- ✓ Open the Data section of SHS 2005
- ✓ Go to "Analysis" on the menu bar and select the "Correlation matrix" program
- ✓ In the Variables to correlate variable boxes copy the variables: "f001 – Food"; "g001 – Shelter"; **"**j001 – Clothing"; and "l201 – Personal care".
- ✓ Select "Run Correlations"
- ✓ The completed cross tab should appear as follows.

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Correlate	**f001**	Food	1.00-64,000.00		1
Correlate	**g001**	Shelter	.00-166,360.00		1
Correlate	**j001**	Clothing	.00-69,300.00		1

Correlate	**l201**	Personal care	.00-11,700.00		1
Weight	**weight**	Weight at household level	10-8479		1

Correlation Matrix

	f001	**g001**	**j001**	**l201**
f001	1.00	.42	.54	.54
g001	.42	1.00	.40	.39
j001	.54	.40	1.00	.65
l201	.54	.39	.65	1.00

All of the correlations are moderate, ranging between 0.4-0.54, except for the relationship between clothing and personal care expenditures which is higher at 0.65.

Basic Data Manipulation

The UT/DLS also several features that allow you to create new variables as well as download the data from the UT/DLS for use in Excel, SPSS, STATA and other statistical packages.

1. Learning Objectives

This section focuses on different methods used to manipulate data using the UT/DLS service. Upon completion of this tutorial you will be able to:

- Recode survey variables
- Compute new variables using existing survey variables
- Calculate dummy variables in order to perform regression analysis
- Download a data subset

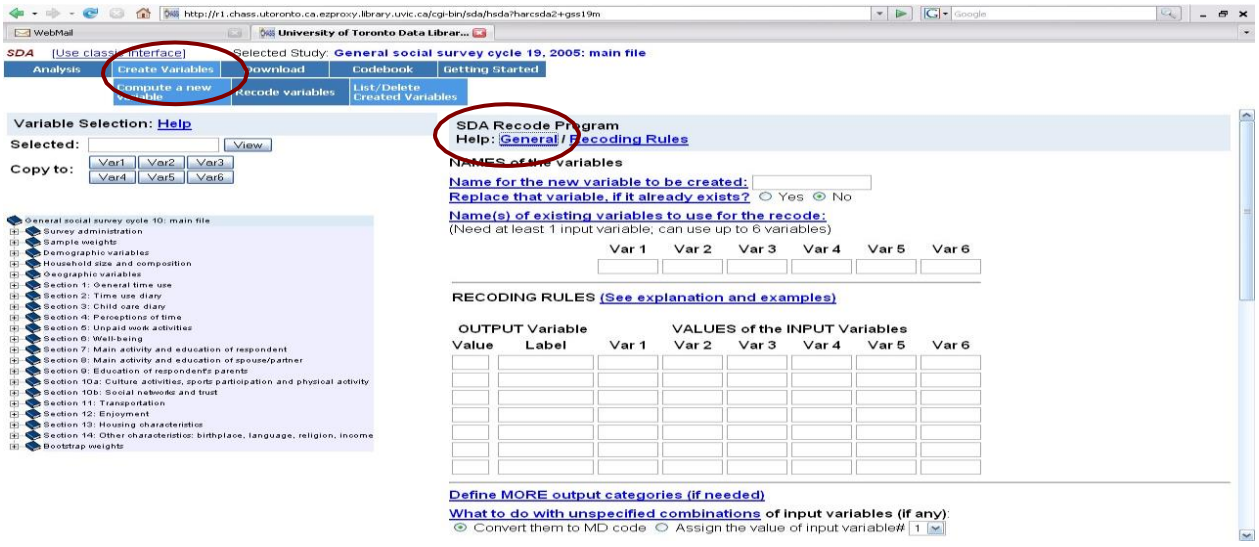
2. Creation of New Variables

Data analysis often requires us to manipulate variables. The UT/DLS service allows you to recode existing variables, compute a new variable by applying operators to existing variables, and create dummy variables for the purposes of performing regression analysis.

2.1 Recoding Variables

The SDA Recode Program allows you to recode one or more existing *numeric* variables into a new UT/DLS variable.

- ✓ Open the SDA Recode Program by highlighting "Create Variables" on the main UT/DLS toolbar and selecting "Recode variables".
- ✓ Read through the SDA Recode Help contents by left clicking "General" at the top of the program



Example: Recoding Data from GSS Cycle 19, 2005

In the General Social Survey on Time Use (GSS Cycle 19) there is a variable "brthprvc – Province of Birth of Respondent" located under variable section 14: Other characteristics: birthplace, language, religion, income. We will recode this variable to reflect *Region of Birth* (Maritimes, Prairies etc.).

- ✓ In "Name for the new variable to be created" type a code for the new variable, Region of Birth e.g. brthreg
- ✓ In the "Var 1" box below "Name(s) of existing variables to use for the recode" enter the name of the variable to be recoded: brthprvc ***or*** select the variable using the Variable Selection tool and copy it to "Var 1".
- ✓ Under OUTPUT Variable enter under the Label column the names Canada's Regions: Maritimes, Quebec, Ontario, Prairies, British Columbia, and Territories.
- ✓ In the Value Column assign values to the new variable labels e.g. 1,2,3,4,5,6
- ✓ Under "VALUES of the INPUT Variables" column Var 1, enter the values of the variable brthprvc that correspond to the new output variable labels (Maritimes = 1-4, Quebec = 5, Ontario = 6 etc.). Consult the codebook to determine the corresponding input variable codes.

brthprvc **Province of birth of R.**

Text of this Question or Item

```
Coverage: All respondents.
Derived from BPR_Q10 and BPR_Q20.
Weight variable: WGHT_PER
```

Percent	N	Valu	Label
6.2	1,199	**1**	Newfoundland and Labrador
2.6	494	**2**	Prince Edward Island
5.2	1,006	**3**	Nova Scotia
5.5	1,069	**4**	New Brunswick
18.9	3,637	**5**	Quebec
21.2	4,083	**6**	Ontario
5.2	1,000	**7**	Manitoba
6.3	1,219	**8**	Saskatchewan
6.8	1,318	**9**	Alberta
6.0	1,147	**10**	British Columbia
0.2	32	**11**	Includes Yukon, Northwest Territories and Nunavut
15.9	3,069	**12**	Countries outside of Canada
	291	**98**	Not stated
	33	**99**	Don't know
100.0	**19,597**		**Total**

Properties

- ✓ Apply desired "OPTIONAL Specification for the New Variable", (i.e. variable label, description etc.)
- ✓ Select "Start Recoding"

The following output will be produced:

UT/DLS 3.1: Recode

General social survey cycle 10: main file

Created Sep 11, 2007 (Tue 02:31 AM EDT)

Variables

Role	Name	Label	Range	MD	Dataset
Output	**brthreg**	Region of Birth	1-6		2
Input	**brthprvc**	Province of birth of R.	1-12	98-*	1

Recode rules

```
Input1:  brthprvc   label: Province of birth of R.

  Output       Input1
       1          1-4
       2            5
       3            6
       4          7-9
       5           10
       6           11
```

Description of the derived variable

brthreg **Region of Birth**

Percent	N	Value	Label
23.3	3,768	**1**	Maritimes
22.4	3,637	**2**	Quebec
25.2	4,083	**3**	Ontario
21.8	3,537	**4**	Prairies
7.1	1,147	**5**	British Columbia
0.2	32	**6**	Territories
	3,393	**<STRONG>.</STRONG>**	(No Data)
100.0	**19,597**		**Total**

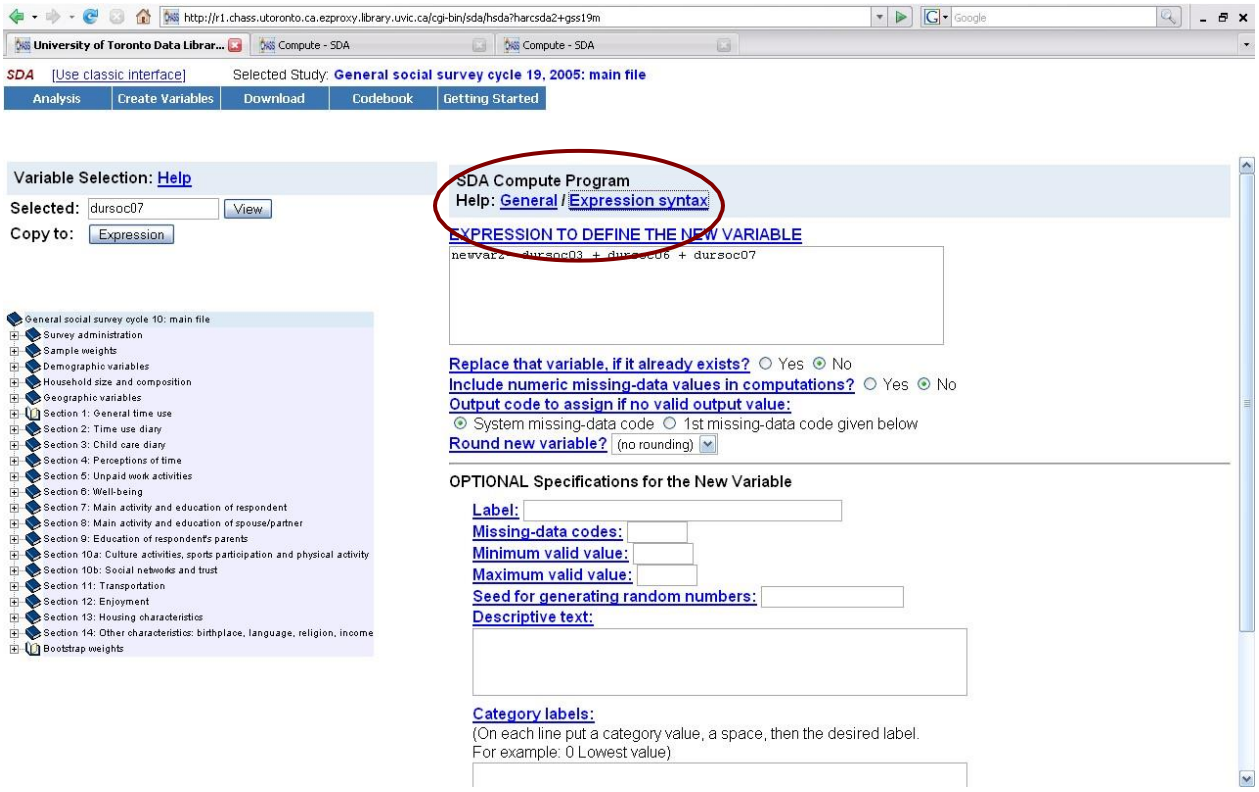
Allocation of cases

Valid cases in new variable	16,204
Cases set to missing-data code	3,393
Total cases	*19,597*

2.2 Computing New Variables

This UT/DLS program creates a new UT/DLS variable as a result of a computation based on one or more existing *numeric* variables.

- ✓ Open the SDA Compute Program by highlighting "Create a Variable" on the main UT/DLS toolbar and selecting "Compute a new variable".
- ✓ Review the General Help and Expression Syntax options to familiarize yourself with the process of computing new variables.



- ✓ Using the General Social Survey on Time Use (Cycle 19) we will compute new variables using two of the more common expression types employed, arithmetic operators and if / else / else if statements.

Example: Computing a new variable using arithmetic operators

We want to create a new variable that reflects the total time *in hours* the respondent spent with their spouse/partner.

- ✓ Using the Variable Selection Tool left click "Section 2: time use diary"
- ✓ Select variable heading "duration by social contact"
- ✓ The variable of interest is: "dursoc02 - Ttl duration (mins.)-social contact - spouse/partner".
- ✓ We will use the variable name "dursocsphr" (total duration of social contact with spouse/partner in hours) to indicate the new variable

- ✓ In the expression box type “dursocsphr = dursoc02/60”
- ✓ In “OPTIONAL Specifications for the New Variable” type beside “Label” type “Total duration of social contact with spouse/partner in hours”.
- ✓ Select “ Start computing”

The following output will be produced:

Variables

Role	Name	Label	Range	MD	Dataset
Output	**dursocsphr**	total duration of social contact with spouse/partner in hours	.0000-24.0000		2
Input	**dursoc02**	Ttl duration (mins.)-social contact - spouse/partner	0-1440		1

Expression used to create the new variable

```
dursocsphr = dursoc02/60
```

Description of the derived variable

```
dursocsphr          total duration of social contact with spouse/partner in
hours
Mean = 2.99842              Std Dev = 4.22595
```

The new variable is now available for use in the data analysis programs. Using the SDA Frequency/Crosstabulation Program determine the frequency distribution of this new variable

- ✓ Open the SDA Frequencies/Crosstabulation Program
- ✓ In the row variable enter “**dursocsphr**”
- ✓ Select “Run the Table”

You will notice from the frequency info produced that the variable takes on so many different values that no discernable pattern can be deduced. Using the If / Else If / Else syntax and logical operators to use with If / Else if we can compute a new variable with fewer potential values so that we can start to discern patterns in the data.

Example: If / Else If / Else syntax and logical operators to use with If / Else

Building on the previous example and utilizing If / Else If / Else syntax and logical operators to use with If / Else we will create a new variable that groups hours of social contact spent with spouse and partner into the following categories:

No time
< 1 hour
1-3 hours
3-5 hours
5-10 hours
10-20 hours
>20 hours

The new variable will be called “scsphr” – social contact spouse/partner hours.

- ✓ Open the SDA “Compute a new variable” Program
- ✓ In the Expression box type the following:

```
If (dursocsphr eq 0) scsphr = 0
else if (dursocsphr gt 0 AND dursocsphr lt 1) scsphr = 1
else if (dursocsphr ge 1 AND dursocsphr lt 3) scsphr = 2
else if (dursocsphr ge 3 AND dursocsphr lt 5) scsphr = 3
else if (dursocsphr ge 5 AND dursocsphr lt 10) scsphr = 4
else if (dursocsphr ge 10 AND dursocsphr lt 20) scsphr = 5
else scsphr = 6
```

- ✓ In “OPTIONAL Specifications for the New Variable” type beside “Label” type “Hours social contact with spouse/partner - 6 categories”.
- ✓ In the “Category labels” box type the following:

0 No time
1 < 1 hour
2 1-3 hours
3 3-5 hours
4 5-10 hours
5 10-20 hours
6 >20 hours

- ✓ Select “Start computing”

The following output will be produced:

Variables

Role	Name	Label	Range	MD	Dataset
Output	**scsphr**	Hours social contact with spouse - 6 categories	0-6		2
Input	**dursocsphr**	total duration of social contact with spouse/partner in hours	.0000-24.0000		2

Expression used to create the new variable

```
If (dursocsphr eq 0) scsphr = 0
else if (dursocsphr gt 0 AND dursocsphr lt 1) scsphr = 1
else if (dursocsphr ge 1 AND dursocsphr lt 3) scsphr = 2
else if (dursocsphr ge 3 AND dursocsphr lt 5) scsphr = 3
else if (dursocsphr ge 5 AND dursocsphr lt 10) scsphr = 4
else if (dursocsphr ge 10 AND dursocsphr lt 20) scsphr = 5
else scsphr = 6
```

Description of the derived variable

scsphr Hours social contact with spouse - 6 categories

Percent	N	Value	Label
49.6	9,729	**0**	No time
3.6	701	**1**	< 1 hour
10.7	2,105	**2**	1-3 hours
11.2	2,189	**3**	3-5 hours
14.6	2,870	**4**	5-10 hours
10.2	1,998	**5**	10-20 hours
0.0	5	**6**	>20 hours

100.0	**19,597**	**Total**

Mean = 1.7 **Std Dev** = 1.9

The new variable is now available for use in the data analysis programs. Using the SDA Frequency/Crosstabulation Program determine the frequency distribution of this new variable

- ✓ Open the SDA Frequencies/Crosstabulation Program
- ✓ In the row variable enter "**scsphr**"
- ✓ In the *CHART OPTIONS* section select "Bar Chart" beside "Type of Chart"
- ✓ Select "Run the Table"

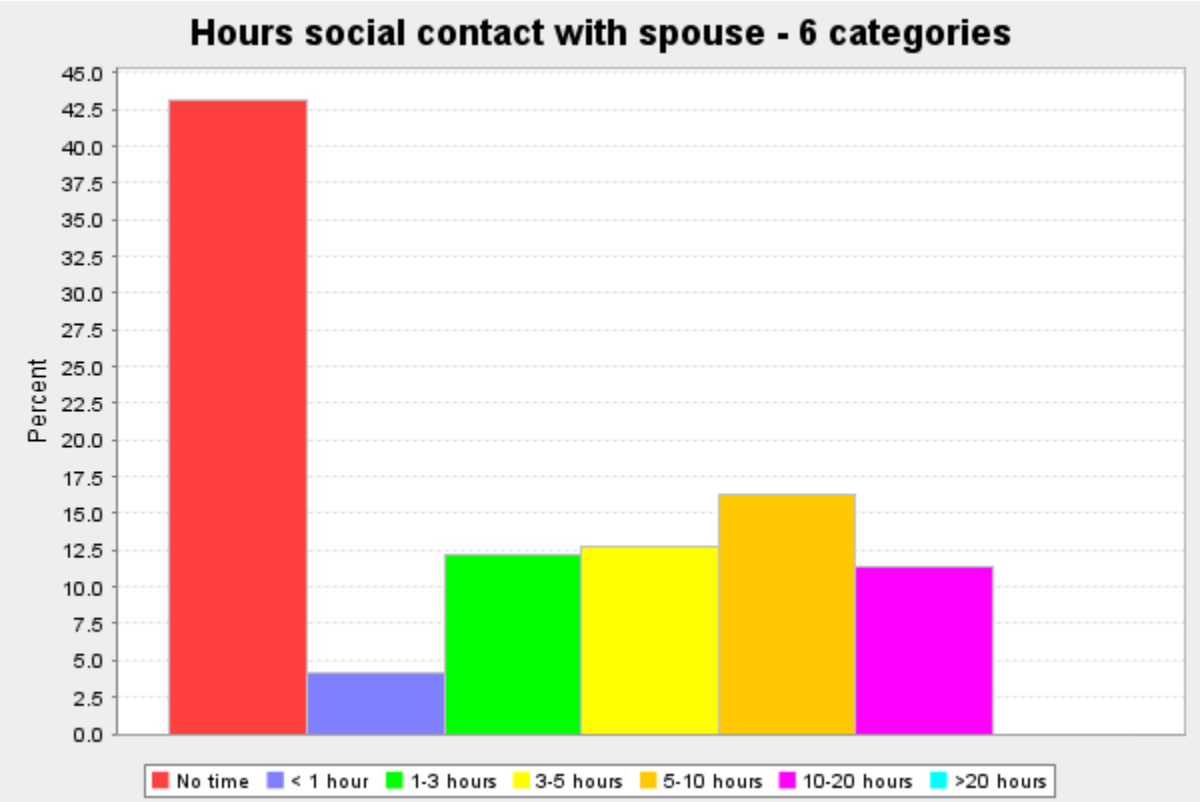
The following frequency distribution and bar chart will be produced:

Variables

Role	Name	Label	Range	MD	Dataset
Row	**scsphr**	Hours social contact with spouse - 6 categories	0-6		2
Weight	**wght_per**	Person weight.	35.7909-10,125.3724		1

Frequency Distribution

Cells contain: -Column percent -N of cases		**Distribution**
scsphr	0: No time	**43.1** 11,253,769
	1: < 1 hour	**4.2** 1,085,542
	2: 1-3 hours	**12.2** 3,190,881
	3: 3-5 hours	**12.8** 3,334,316
	4: 5-10 hours	**16.3** 4,252,628
	5: 10-20 hours	**11.4** 2,969,632
	6: >20 hours	**.0** 9,050
	COL TOTAL	***100.0*** *26,095,819*



2.3 Creating Dichotomies and Dummy Variables

Dichotomies and dummy variables are used to convert variable values so that they can be used in performing regression analysis. A dichotomy is created when a variable is simply coded as either 0 or 1 (e.g. male(0) or female(1); foreign(0) or domestic(1). A system of dummy variables is used to create dichotomies when a non-interval variable has more than two categories. Creating dichotomies and dummy variable involves recoding existing variables and/or creating new variables.

Example: Recode variable "Sex – Sex of respondent" into a dichotomy

- ✓ A demographic variable that describes the sex of respondents is included in all surveys of interest however for this example we will use the GSS Cycle 19
- ✓ Open the Data section of GSS Cycle 19
- ✓ Open the SDA "Recode variables" Program
- ✓ Using the Variable Selection tool, copy the variable "sex – Sex of R" into the "Var1" box on the recode program
- ✓ Name the new variable to be created "sexD" to indicate that it is the variable Sex recoded as a dichotomy
- ✓ In the "Recoding Rules" section enter the dichotomized variable values (0 and 1) into the "OUTPUT Variables" Values boxes.
- ✓ In the "INPUT Variables" section enter "Male and Female" into the Label boxes beside the OUTPUT Variable Values 0 and 1 respectively.
- ✓ Enter the corresponding INPUT Variables values for the two labels into the VAR1 boxes (1 for Male; 2 for Female).

- ✓ In the "OPTIONAL Specifications for the New Variable" Section enter "Sex of Respondent – Dichotimized" in the Label box.
- ✓ Select "Start Recoding"

The following output will be produced:

UT/DLS 3.1: Recode

General social survey cycle 10: main file

Created Oct 09, 2007 (Tue 03:56 PM EDT)

Variables

Role	Name	Label	Range	MD	Dataset
Output	**sexd**	Sex of Respondent – Dichotimized	0-1		2
Input	**sex**	Sex of R.	1-2		1

Recode rules

```
Input1:  sex    label: Sex of R.

  Output       Input1
       0            1
       1            2
```

Description of the derived variable

sexd Sex of Respondent – Dichotimized

Percent	N	Value	Label
44.0	8,621	**0**	Male
56.0	10,976	**1**	Female
100.0	**19,597**		**Total**

Allocation of cases

Valid cases in new variable	19,597
Cases set to missing-data code	0
Total cases	*19,597*

Data sets

1	/gss/gss19/gss19m
2	/dli/gss/gss19m

Example – Creating dummy variables from non-interval variables with multiple categories

Non-interval variables that have multiple categories can be incorporated into a multiple regression by creating a system of dummy variables. Using the Canadian Community Health Survey (CCHS) Cycle 3.1, 2005, we will create a series of dummy variables to from the variable "eduedh04 - Highest level/edu. - HH 4 levels - (D)".

- ✓ Open the SDA "Compute a new variable" Program
- ✓ Using the Variable Selection Tool open the variable heading category "EDU Education" and select the variable "eduedh04 - Highest level/edu. - HH 4 levels - (D)".
- ✓ Select view to see the variable values (the following information will be created in a new window:

eduedh04 Highest level/edu. - HH 4 levels - (D)

Percent	N	Value	Label
14.4	17,410	**1**	< THAN SECONDARY
12.2	14,811	**2**	SECONDARY GRAD.
6.1	7,420	**3**	OTHER POST-SEC.
67.3	81,493	**4**	POST-SEC. GRAD.
	11,087	**9**	NOT STATED
100.0	**132,221**		**Total**

Properties	
Data type:	numeric
Missing-data codes:	6-9
Mean:	3.26
Std Dev:	1.14
Record/column:	1/1576

- ✓ When creating a system of dummy variables the variable categories must be recoded as dichotomies.
- ✓ When using dummy variables it is important to remember to create one fewer dummy variables than there are categories in the non-interval variable being represented. Consequently, for this example we will be creating 3 dummy variables, each represented by the following dichotomies:

D1 = a dummy variable scored 1 if highest level of household education = 1(< than secondary) and 0 otherwise

D2 = a dummy variable scored 1 if highest level of household education = 2(secondary grad) and 0 otherwise

D3 = a dummy variable scored 1 if highest level of household education = 3(other post-sec) and 0 otherwise

The SDA Compute Program only allows us to compute one new variable at a time, therefore we will start with creating dummy variable D1.

- ✓ In the expression box type the following:

If (eduedh04 eq 1) D1 = 1
Else D1 = 0

- ✓ In the Label box type – Dummy variable education < secondary
- ✓ Select "Start computing".

The following output will be produced:

Variables

Role	Name	Label	Range	MD	Dataset
Output	**d1**	Dummy variable education < secondary	0-1		2
Input	**eduedh04**	Highest level/edu. - HH 4 levels - (D)	1-4	6-9	1

Expression used to create the new variable

```
If (eduedh04 eq 1) D1 = 1
Else D1 = 0
```

Description of the derived variable

d1 Dummy variable education < secondary

Percent	N	Value	Label
85.6	103,724	**0**	
14.4	17,410	**1**	
	11,087	**<STRONG>.</STRONG>**	(No Data)
100.0	**132,221**		**Total**

Mean = .1 **Std Dev** = .4

- ✓ Repeat the above steps for the other dummy variables D2 and D3 using the following expressions:

D2 – dummy variable education = secondary grad

If (eduedh04 eq 2) D2 = 1
Else D2 = 0

Output:

Variables

Role	Name	Label	Range	MD	Dataset
Output	**d2**	Dummy variable education = secondary grad	0-1		2
Input	**eduedh04**	Highest level/edu. - HH 4 levels - (D)	1-4	6-9	1

Expression used to create the new variable

```
If (eduedh04 eq 2) D2 = 1
Else D2 = 0
```

Description of the derived variable

d2 Dummy variable education = secondary grad

Percent	N	Value	Label
87.8	106,323	**0**	
12.2	14,811	**1**	
	11,087	**<STRONG>.</STRONG>**	(No Data)
100.0	**132,221**		**Total**

Mean = .1 **Std Dev** = .3

D3 – dummy variable education = other post secondary

If (eduedh04 eq 3) D3 = 1
Else D3 = 0

Output:

Variables

Role	Name	Label	Range	MD	Dataset
Output	**d3**	Dummy variable education = other post sec	0-1		2
Input	**eduedh04**	Highest level/edu. - HH 4 levels - (D)	1-4	6-9	1

Expression used to create the new variable

```
If (eduedh04 eq 3) D3 = 1
Else D3 = 0
```

Description of the derived variable

d3 Dummy variable education = other post sec

Percent	N	Value	Label
93.9	113,714	**0**	
6.1	7,420	**1**	
	11,087	**<STRONG>.</STRONG>**	(No Data)
100.0	**132,221**		**Total**

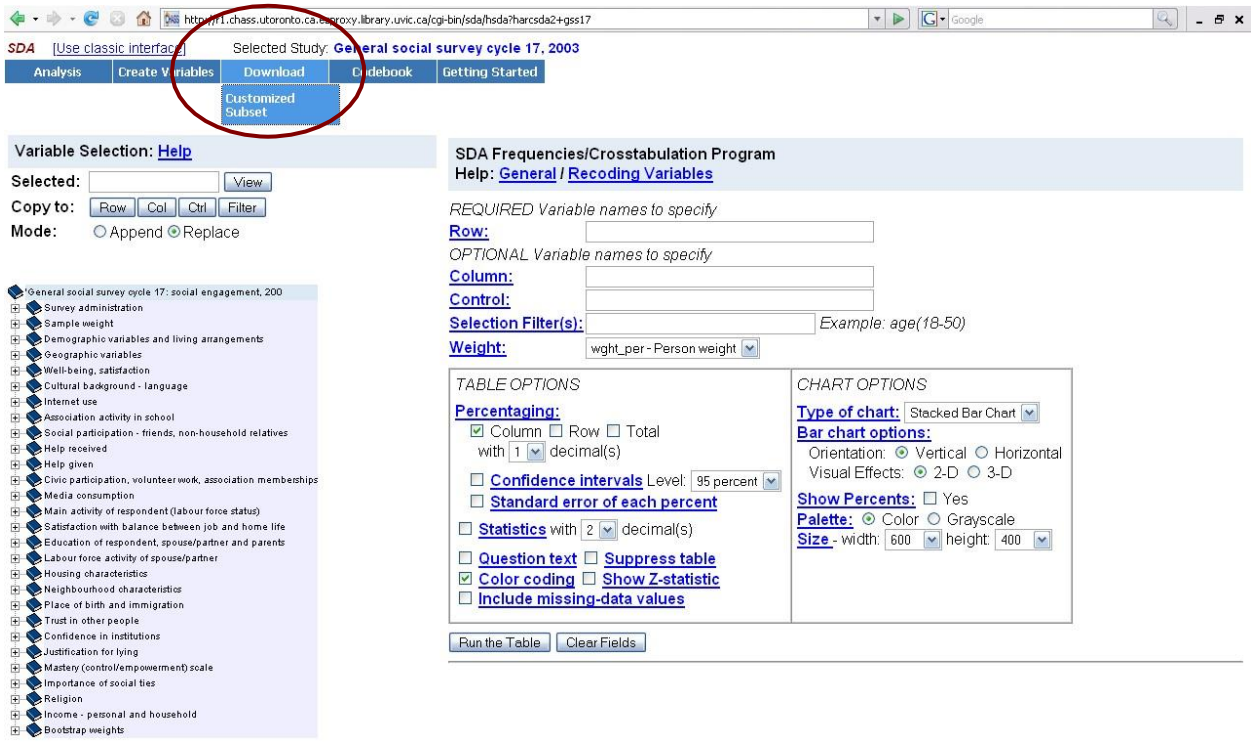
Mean = .1 **Std Dev** = .2

3. Downloading Data Subset to Excel

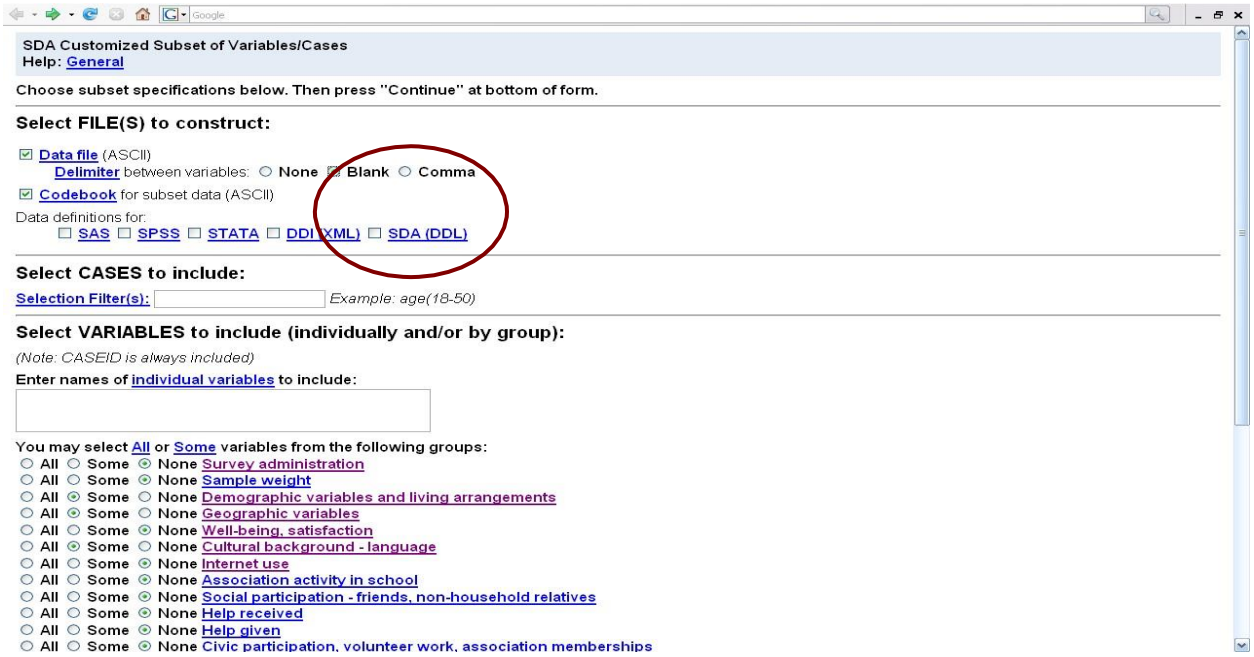
You may often want to download raw data from the UT/DLS to Excel (or SAS, SPSS, STATA) in order to graph results or perform other statistical analysis.

Example: Download GSS Cycle 17 data to Excel

- ✓ Left click on the "Data" link next the survey.
- ✓ Highlight "Download" on the toolbar at the top of the page and select the "Customized Subset" option.

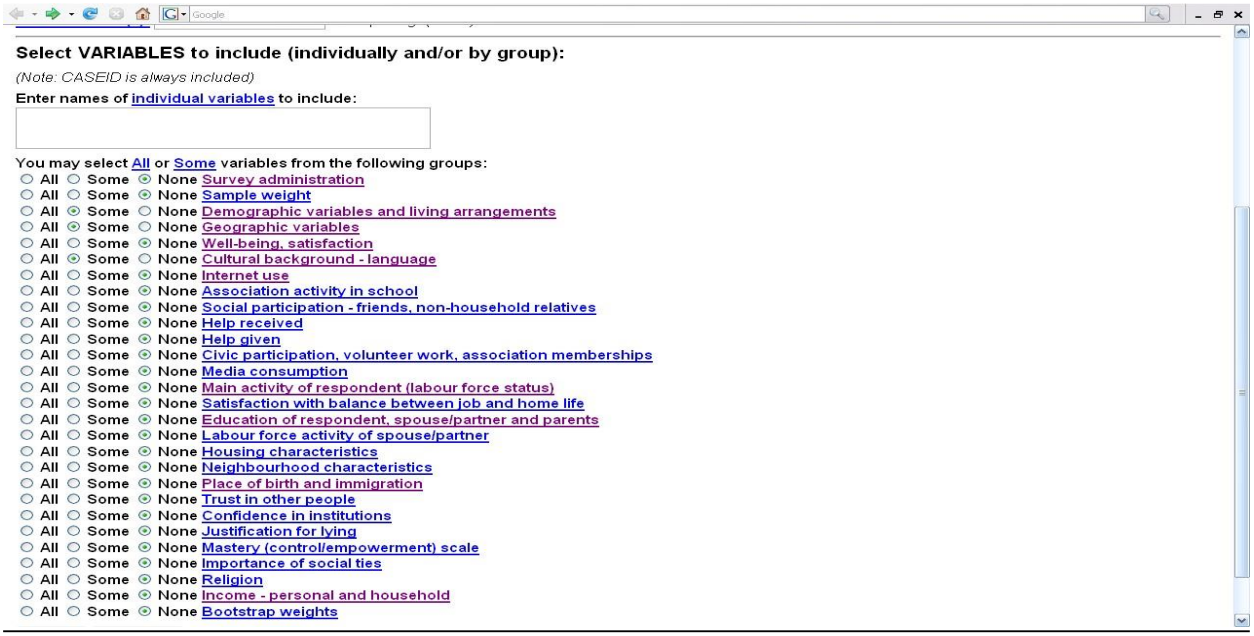


- ✓ The UT/DLS Customized Subset of Variables/Cases program will open in a new window.
- ✓ Select either "Blank" or "Comma" as a Delimiter between variables. **<u>If a delimiter is not selected the variables downloaded will not be separated and then cannot be distributed across spreadsheet columns.</u>**
- ✓

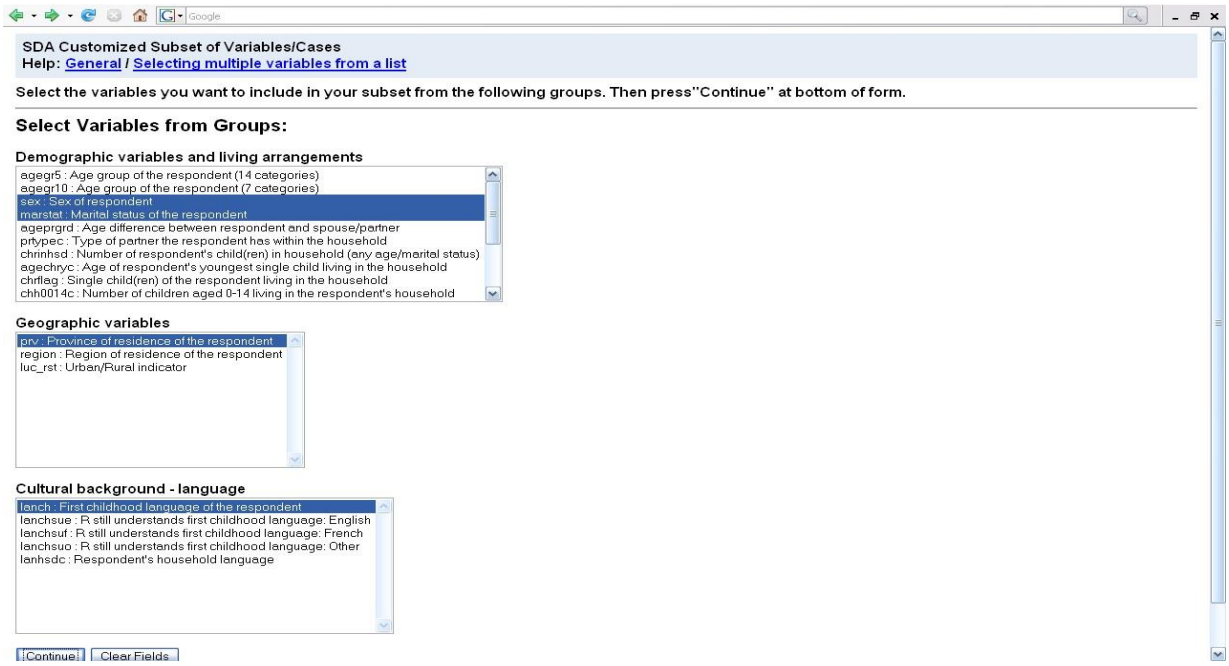


- ✓ If you are only interested in a subgroup of the sample (e.g. women) apply the appropriate selection filter.

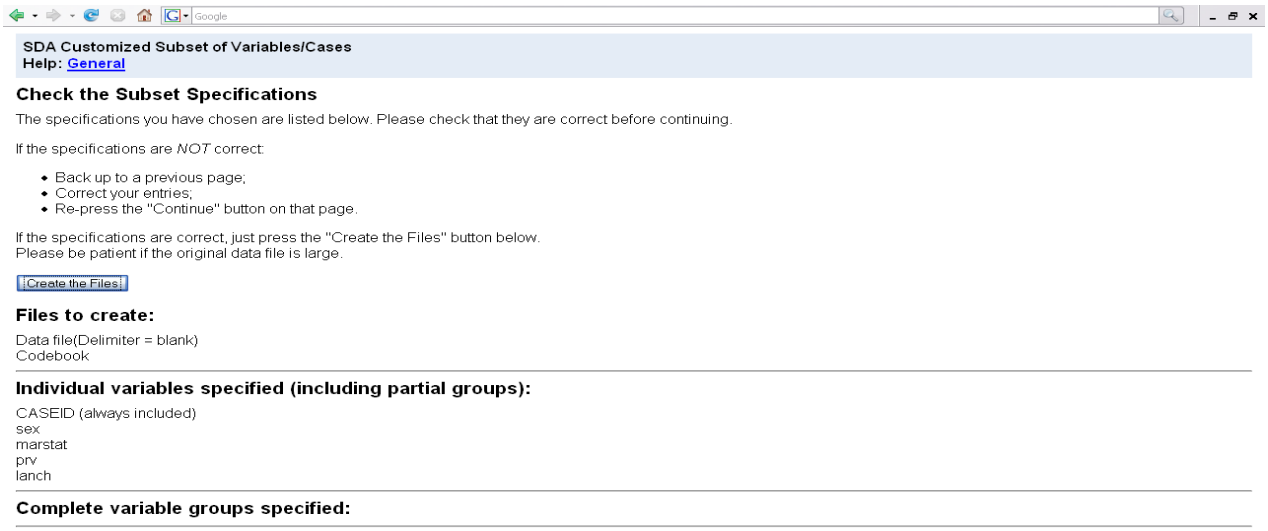
- ✓ Select the variables you wish to download. These can be entered individually by entering the desired variable codes from the codebook or by groups.
- ✓ To select variables by group, select "Some" or "All" next to the desired variable categories (e.g. Demographic variables and living arrangements, Geographic variables).
- ✓ Select "Continue" at the bottom of the page.



- ✓ Select the desired variables to download from the group lists by clicking to highlight them. To select more than one variable from an individual group hold down the "Ctrl" button and using the mouse click on the desired variables.
- ✓ When all of the desired variables are selected click "Continue".



- ✓ Left click “Create the Files” when you are satisfied with the Individual variables specified



SDA Customized Subset of Variables/Cases
Help: General

Check the Subset Specifications

The specifications you have chosen are listed below. Please check that they are correct before continuing.

If the specifications are *NOT* correct:

- Back up to a previous page;
- Correct your entries;
- Re-press the "Continue" button on that page.

If the specifications are correct, just press the "Create the Files" button below.
Please be patient if the original data file is large.

[Create the Files]

Files to create:

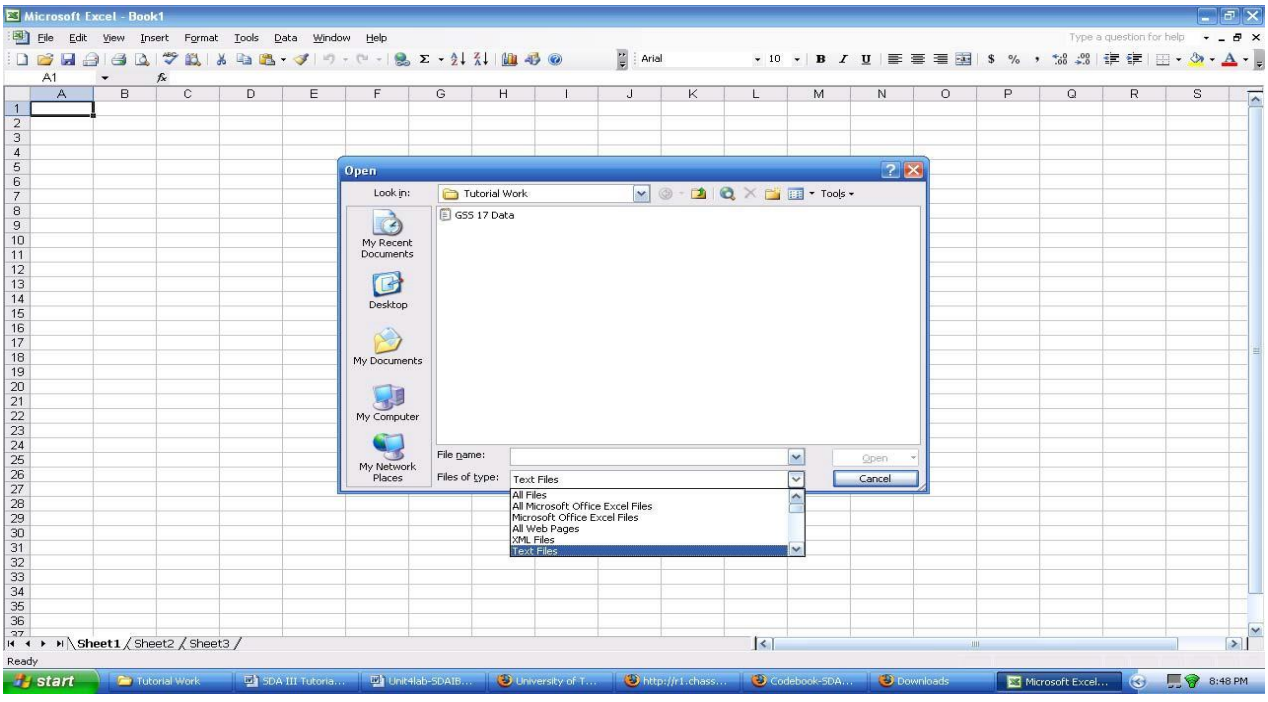
Data file(Delimiter = blank)
Codebook

Individual variables specified (including partial groups):

CASEID (always included)
sex
marstat
prv
lanch

Complete variable groups specified:

- ✓ To view or save the file follow the instructions on the page. The file will be saved as a text document
- ✓ Open Excel and select the “File/Open” option
- ✓ Select the Text data file downloaded from the UT/DLS



- ✓ On the Text Import Window under “Original data type” select “Fixed width”. Select Next.

- ✓ Confirm the desired data preview and adjust accordingly.
- ✓ Select Next.
- ✓ Confirm the data format and select Finish.
- ✓ Label the data columns accordingly. Variables will be downloaded in the order indicated in the downloaded codebook

4. Conclusion

This tutorial covers the basic material for data manipulation using the UT/DLS service. To ensure that you understand the basic material presented in this section, please work through the following exercises.

5. Exercises

a) Using the Canadian Community Health Survey (CCHS) Cycle 3.1, 2005, recode the variable "dhhegage – Age – (G)" to create a new variable "gen – Generations" that reflects the following generational labels commonly used to distinguish different demographic segments of the workforce:

Mature: ≥ 60
Boomers: 40 - 59
Generation X: 25- 39
Millenials: < 25

b) Using the Canadian Community Health Survey (CCHS) Cycle 3.1, 2005, create a system of dummy variables from the variable "lbsedwss - Working status last week - 4 grps - (D)".

6. Answers

a)

- ✓ Open the SDA Recode Program
- ✓ Name the new variable to be created "gen" indicating Generations

- ✓ Using the Variable Selection tool copy the variable "dhhegage – Age – (G)" into "Var 1" of "Name(s) of existing variables to use for the recode"
- ✓ Enter the following into the OUTPUT Variables Labels Column: Millenials; Generation X, Boomers, and Mature.
- ✓ Assign values to the OUTPUT variable labels
- ✓ Consult the codebook to determine the variable values for the new labels.

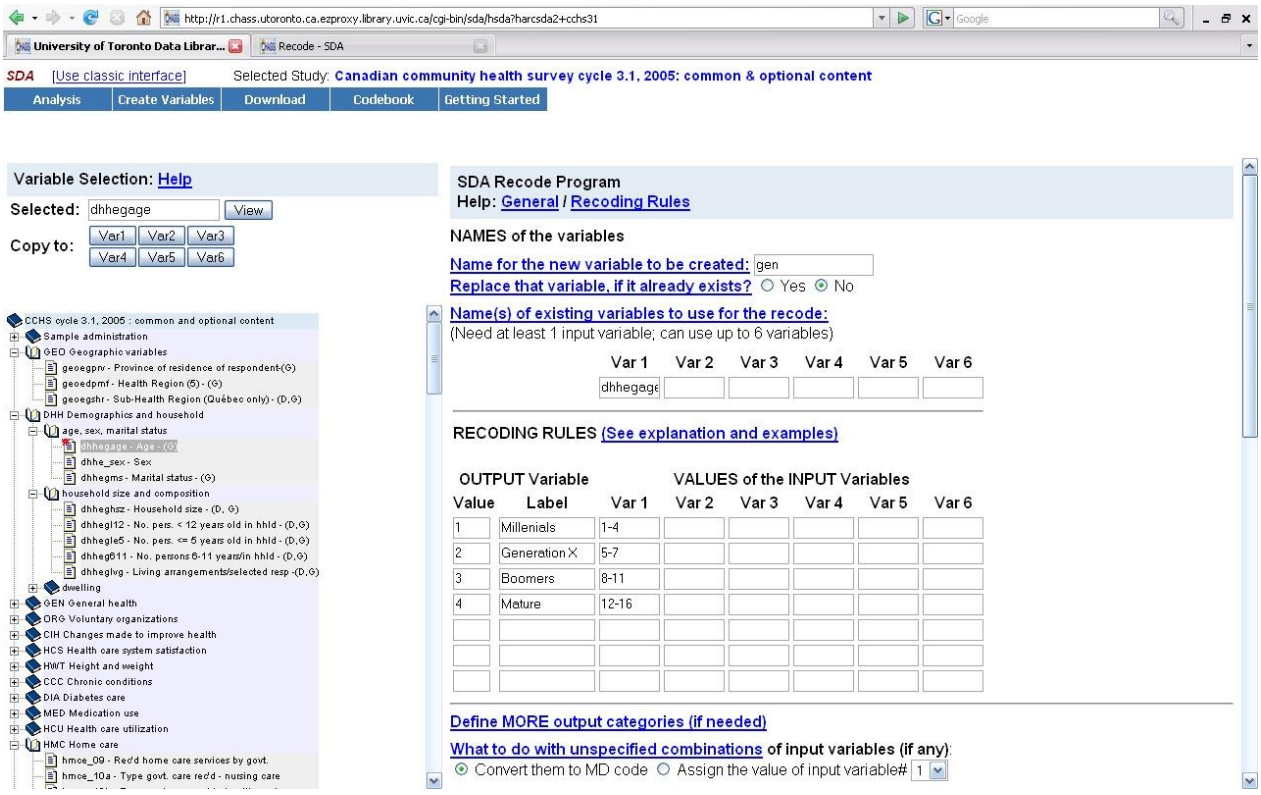
dhhegage **Age - (G)**

Percent	N	Value	Label
4.7	6,172	**1**	12 TO 14 YEARS
4.6	6,145	**2**	15 TO 17 YEARS
3.0	3,989	**3**	18 TO 19 YEARS
5.9	7,740	**4**	20 TO 24 YEARS
7.0	9,227	**5**	25 TO 29 YEARS
7.8	10,252	**6**	30 TO 34 YEARS
7.6	10,058	**7**	35 TO 39 YEARS
8.4	11,172	**8**	40 TO 44 YEARS
6.9	9,143	**9**	45 TO 49 YEARS
7.8	10,296	**10**	50 TO 54 YEARS
8.1	10,645	**11**	55 TO 59 YEARS
7.0	9,268	**12**	60 TO 64 YEARS
5.9	7,846	**13**	65 TO 69 YEARS
5.4	7,124	**14**	70 TO 74 YEARS
4.5	5,961	**15**	75 TO 79 YEARS
5.4	7,183	**16**	80 YEARS OR MORE
100.0	**132,221**		**Total**

Properties

Data type: numeric
Record/columns: 1/26-27

- ✓ Enter the following values under the INPUT variable Var 1 Column for the corresponding labels:
  Millenials: 1-4
  Generation X: 5-7
  Boomers: 8-11
  Mature: 12-16



- ✓ Enter "Generations" in OPTIONAL "Label Box"
- ✓ Select "Start Recoding"
- ✓ The following output should be produced:

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Output	**gen**	Generations	1-4		2
Input	**dhhegage**	Age - (G)	1-16		1

Recode rules

```
Input1:  dhhegage   label: Age - (G)

  Output      Input1
       1         1-4
       2         5-7
       3        8-11
       4       12-16
```

Description of the derived variable

gen Generations

Percent	N	Value	Label
18.2	24,046	**1**	Millenials
22.3	29,537	**2**	Generation X
31.2	41,256	**3**	Boomers

28.3	37,382	**4**	Mature
100.0	**132,221**		**Total**

b)

- ✓ Open the Data section of the CCHS, Cycle 3.1
- ✓ Open the SDA Compute Program
- ✓ Using the Variable Selection tool expand the variable heading "LBS Labour force activity" and select the variable "lbsedwss - Working status last week - 4 grps - (D)".
- ✓ Select the "View" button to determine the potential values of the variables

lbsedwss **Working status last week - 4 grps - (D)**

Percent	N	Value	Label
59.3	65,653	**1**	AT WORK LAST WK
5.5	6,030	**2**	ABSENT LAST WK
32.4	35,851	**3**	NO JOB LAST WK
2.8	3,096	**4**	UNABLE/PERMANENT
	19,316	**6**	NOT APPLICABLE
	2,275	**9**	NOT STATED
100.0	**132,221**		**Total**

- ✓ Create the following system of Dummy Variables

D1 = a dummy variable scored 1 if work status last week = 1(at work last wk) and 0 otherwise

D2 = a dummy variable scored 1 if work status last week = 2(absent last wk) and 0 otherwise

D3 = a dummy variable scored 1 if work status last week = 3(no job last wk) and 0 otherwise

Dummy variable D1:

- ✓ In the expression box type the following:

If (lbsedwss eq 1) D1 = 1
Else D1 = 0

- ✓ In the Label box type – Dummy variable work status last week – at work
- ✓ In "Category labels" type:

0 = Not at work last week
1 = At work last week

- ✓ Select "Start computing".
- ✓ Output created:

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Output	**d1**	Dummy variable work status last week = at work	0-1		2
Input	**lbsedwss**	Working status last week - 4 grps - (D)	1-4	6-9	1

Expression used to create the new variable

```
If (lbsedwss eq 1) D1 = 1
Else D1 = 0
```

Description of the derived variable

d1 Dummy variable work status last week = at work

Percent	N	Value	Label
40.7	44,977	**0**	= Not at work last week
59.3	65,653	**1**	= At work last week
	21,591	**<STRONG>.</STRONG>**	(No Data)
100.0	**132,221**		**Total**

Mean = .6 **Std Dev** = .5

Dummy variable D2:

- ✓ In the expression box type the following:

If (lbsedwss eq 2) D2 = 1
Else D2 = 0

- ✓ In the Label box type – Dummy variable work status last week – absent
- ✓ In "Category labels" type:

0 = Not absent from work last week
1 = Absent from work last week

- ✓ Select "Start computing".

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Output	**d2**	Dummy variable work status last week = absent from work	0-1		2
Input	**lbsedwss**	Working status last week - 4 grps - (D)	1-4	6-9	1

Expression used to create the new variable

```
If (lbsedwss eq 2) D2 = 1
Else D2 = 0
```

Description of the derived variable

d2 Dummy variable work status last week = absent from work

Percent	N	Value	Label

94.5	104,600	**0**	= Not absent from work last week
5.5	6,030	**1**	= Absent from work last week
	21,591	**<STRONG>.</STRONG>**	(No Data)
100.0	**132,221**		**Total**

Mean = .1 **Std Dev** = .2

Dummy variable D3:

- ✓ In the expression box type the following:

If (lbsedwss eq 3) D3 = 1
Else D3 = 0

- ✓ In the Label box type – Dummy variable work status last week – no job
- ✓ In "Category labels" type"

0 = Not absent from work last week
1 = Absent from work last week

- ✓ Select "Start computing".

Variables

Role	Name	Label	Range	MD	Dataset
Output	**d3**	Dummy variable work status last week = no job	0-1		2
Input	**lbsedwss**	Working status last week - 4 grps - (D)	1-4	6-9	1

Expression used to create the new variable

```
If (lbsedwss eq 3) D3 = 1
Else D3 = 0
```

Description of the derived variable

d3 **Dummy variable work status last week = no job**

Percent	N	Value	Label
67.6	74,779	**0**	= Job last week
32.4	35,851	**1**	= No job last week
	21,591	**<STRONG>.</STRONG>**	(No Data)
100.0	**132,221**		**Total**

Mean = .3 **Std Dev** = .5