Epi Info 7 Course

Analyzing Global Youth Tobacco Survey Data with Epi Info 7

Table of Contents

Unweighted. A point estimate (percentage) that was NOT calculated using the sampling weights. Unweighted estimates should NOT be used when reporting point estimates (percentages).

Weighted. An estimate (percentage) that was created using the sampling weights.

Common Abbreviations and Acronyms

CRXX	A variable name indicating the Core GYTS questionnaire number
	Examples:
	CR1 = GYTS Core Question number 1
	CR5 = GYTS Core Question number 5

FINALWGT	Final sampling weight variable on your GYTS data set
LCL %	Lower Confidence Limit. Lower bound of 95% Confidence Interval
n	Sample Size
OSH	Office on Smoking and Health at the CDC
PSU	Primary Sampling Unit variable on your GYTS data set
SE	Standard Error
STRATUM	Stratum ID variable on your GYTS data set
95% CI	95% confidence Interval
UCL %	Upper Confidence Limit. Upper bound of 95% Confidence Interval

Part One. Introduction

Mission

Epi Info 7 is a free statistical analysis software produced by the Centers for Disease Control and Prevention which can analyze complex survey designs such as the Global Youth Tobacco Survey.

This course will provide you with basic installation and data analysis training for analyzing complex survey data.

Working through this Course

This course is progressive, and it is important to follow through each part sequentially. Beyond this course, you can study the Epi Info 7 User Guide or the equivalent online Epi Info 7 Help, as well as network with other epidemiologists in your region who use Epi Info 7.

About Epi Info 7

We recommend using Epi Info 7 software because it is free and because it provides a Complex Sample module that produces appropriate standard error estimates and confidence intervals for complex survey designs such as the Global Youth Tobacco Survey (GYTS). Epi Info 7 is a proven tool among many scientists around the world and has a network of users for technical support. The Epi Info 7 User Guide includes contact information for technical support.

There are other statistical packages available including SUDAAN, STATA, SPSS, and SAS which take the complex survey design into account when computing standard errors or confidence intervals. You may use one of these packages; however, unless you have specific training in these other software packages, we recommend using Epi Info 7.

Course Software Version

We are using Epi Info 7 version 7.2.0.1 in this course.

General Information and Technical Support

Because Epi Info 7 is not copyrighted, you can legally install copies of the software on any number of computers; the software is freely distributed on the Internet. To find out more about Epi Info 7 and to access the most recent free information, technical support, and downloads, go to the following URL: http://www.cdc.gov/epiinfo/pc/index.html.

Translations

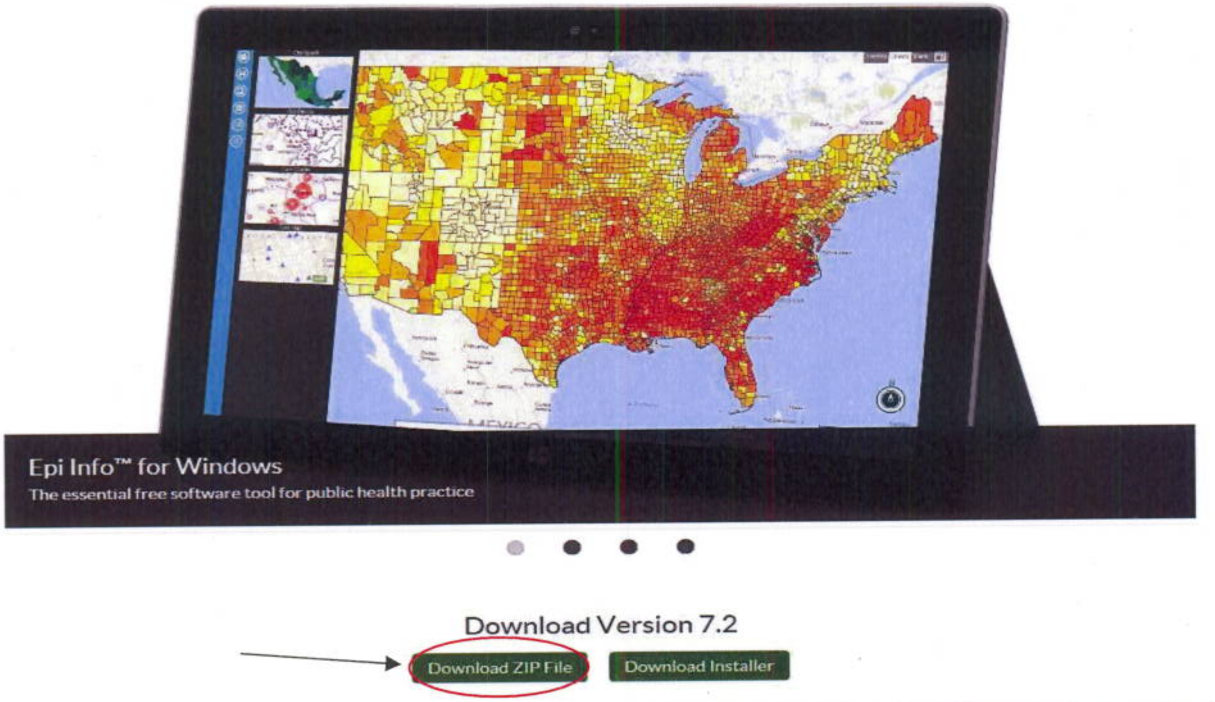
Epi Info 7 (software and manual) is currently available in English, and translation databases are also available in eight languages. These versions are available from specific contacts around the world, and the most recent language and contact information is available at the following URL: http://www.cdc.gov/epiinfo/support/translations.html.

Part Two. Downloading and Installing Epi Info 7

DOWNLOADING Epi Info 7

Epi Info 7 may be downloaded as a “zip” file (recommended) or as a “setup” file. The “zip” file can be downloaded to most user desktops and run without requiring administrative or elevated privileges. The “setup” file requires administrative or elevated privileges during installation.

We are providing the Epi Info 7 “zip” file on a thumb drive, and you may take this thumb drive with you to install the software on other computers. You may also download the “zip” file from the Epi Info 7 website: http://www.cdc.gov/epiinfo/pc/index.html.



Epi Info™ for Windows

The essential free software tool for public health practice

Download Version 7.2

We recommend saving the Epi Info 7 “zip” file to your desktop. Before downloading Epi Info 7, make sure your computer meets the minimum system requirements below.

Required	• Microsoft Windows XP or above • Microsoft .NET Framework 4.0 or above
Recommended	• 256 MB of RAM • 1 GHz processor

INSTALLING Epi Info 7

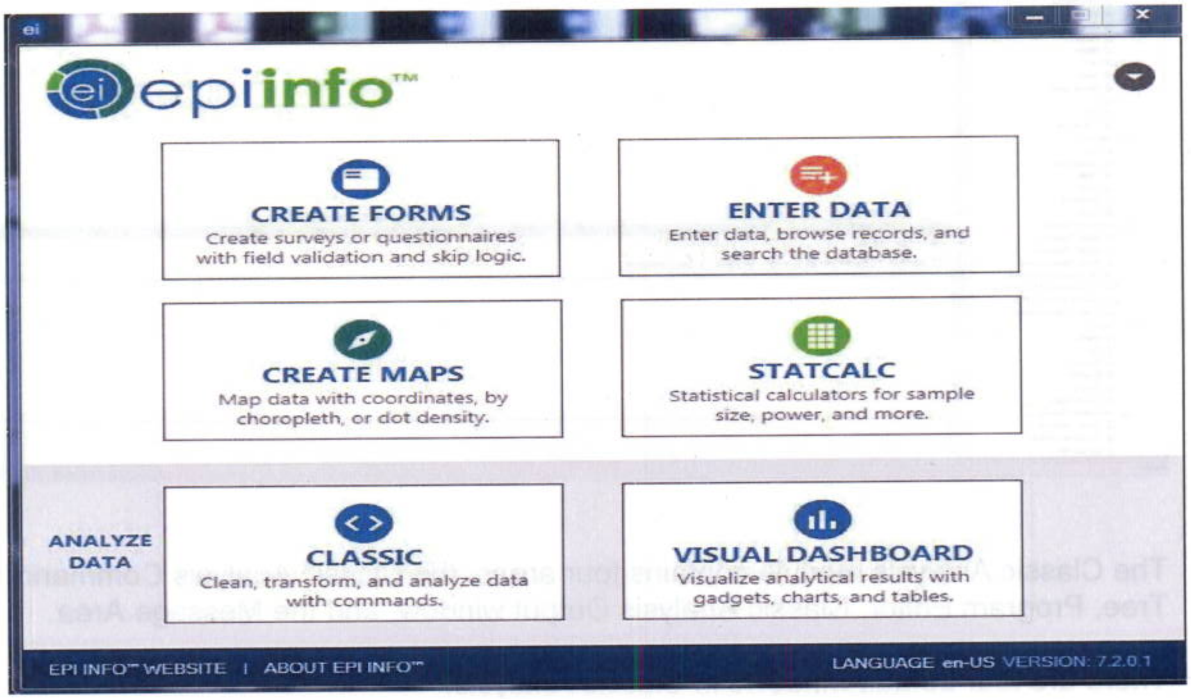
Before installing Epi Info 7, uninstall previous versions of Epi Info. If you have a previous Epi Info version, you can uninstall it using the Control Panel, then the "Add/Remove Programs" selection.

1. After downloading the zip file, go to your desktop and open the zip file.
2. Click on “Extract” at the top to extract the Epi Info 7 folder and .exe file to your desktop.
3. Go back to the desktop and double-click on the “Launch Epi Info 7.exe” icon.

Part Three. Introduction to Analysis

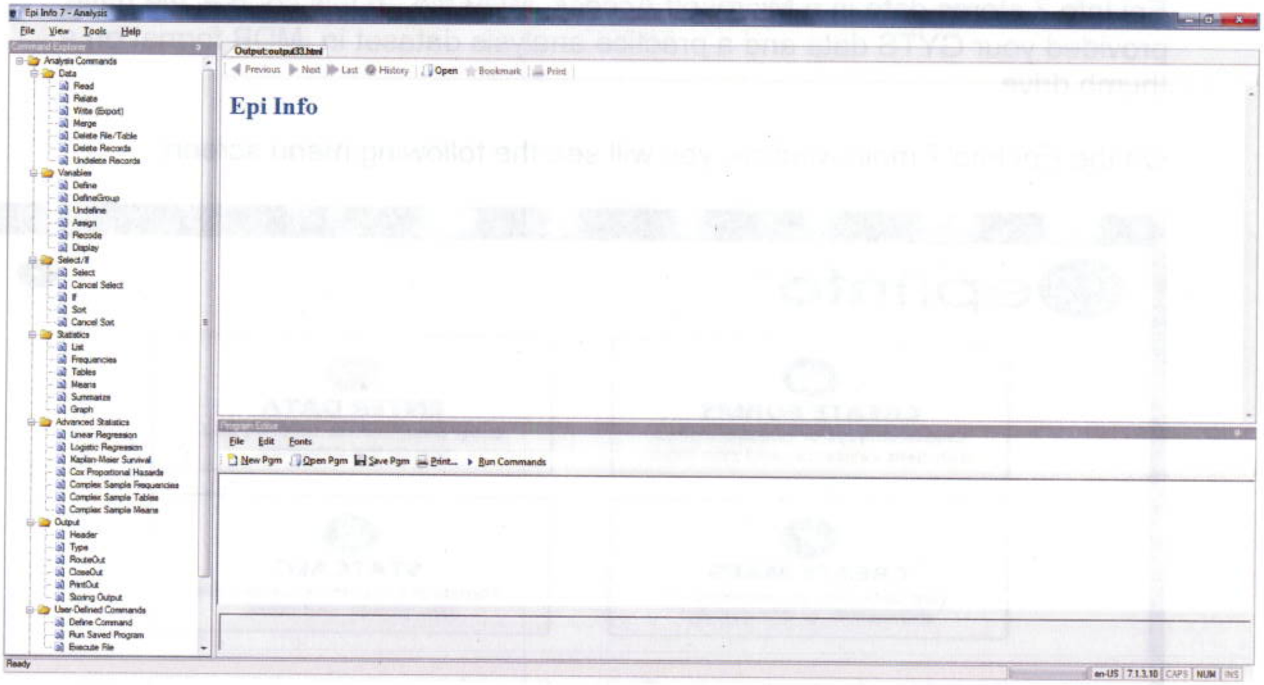
Epi Info 7 stores data in a Microsoft Access .MDB file. In this course, we have provided your GYTS data and a practice analysis dataset in .MDB format on a thumb drive.

On the Epi Info 7 main window, you will see the following menu screen:


ei
epiinfo™
CREATE FORMS
Create surveys or questionnaires with field validation and skip logic.
ENTER DATA
Enter data, browse records, and search the database.
CREATE MAPS
Map data with coordinates, by choropleth, or dot density.
STATCALC
Statistical calculators for sample size, power, and more.
ANALYZE DATA
CLASSIC
Clean, transform, and analyze data with commands.
VISUAL DASHBOARD
Visualize analytical results with gadgets, charts, and tables.
EPI INFO™ WEBSITE | ABOUT EPI INFO™
LANGUAGE en-US VERSION: 7.2.0.1

Epi Info 7 has six modules: Create Forms, Enter Data, Create Maps, StatCalc, Classic Analysis, and Visual Dashboard. To analyze complex GYTS data, we will use Classic Analysis.

Next, click the "Classic" menu button, and you will see a screen display as follows:



The Classic Analysis module contains four areas: the Classic Analysis Command Tree, Program Editor, Classic Analysis Output window, and the Message Area.

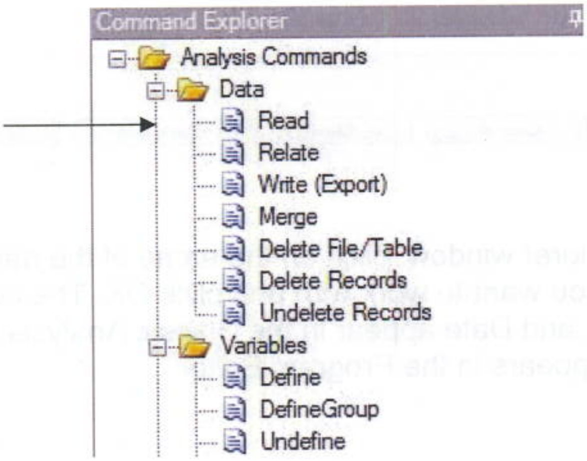
There are four default windows in Classic Analysis:

- The Analysis Commands window on the left contains a list of the commands available in the Analysis module.
- The Output window displays information and output generated from the Analysis Commands.
- The Program Editor window displays the commands and code created using the analysis commands. It is also the place to write and store Epi Info 7 commands.
- The Message window alerts you if any problems occur with any executed commands.

Part Four. Reading and Importing Data

In Epi Info 7, an MDB file as considered a project, and this project may contain several tables (each of which can be thought of as a dataset). The advantage of the MDB format is that your data and tables will all be in one file.

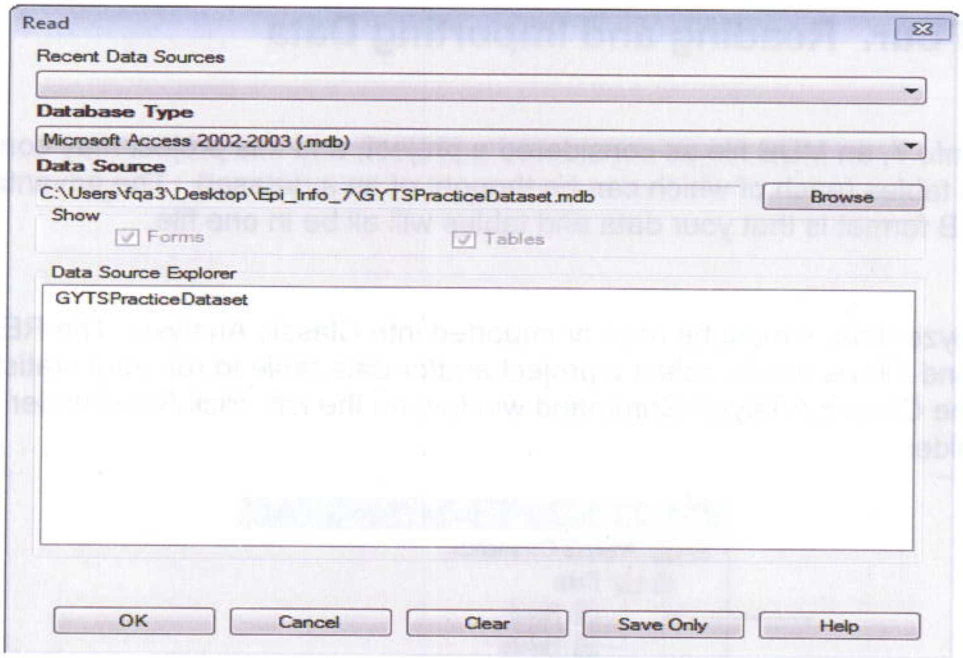
To analyze data, it must be read or imported into Classic Analysis. The READ command allows you to select a project and/or data table to run your statistics. From the Classic Analysis Command window on the left, click Read under the Data folder.



The Recent Data Sources field provides a list of recently-accessed databases in Classic Analysis or Visual Dashboard. By selecting one of the data sources available on the list, you do not need to provide Database Type or Data Source information because they will populate automatically. If you have not worked in Epi Info 7 before, then you can skip the Recent Data Sources field.

The Database Type field indicates the database file to be loaded (.PRJ, .MDB, .XLS). Select Microsoft Access 2002-2003 (.mdb).

The Data Source field indicates the file location/path. Next to the Data Source field, click the Browse button on the right to browse and select the file (GYTS Practice Dataset) to import into Classic Analysis.



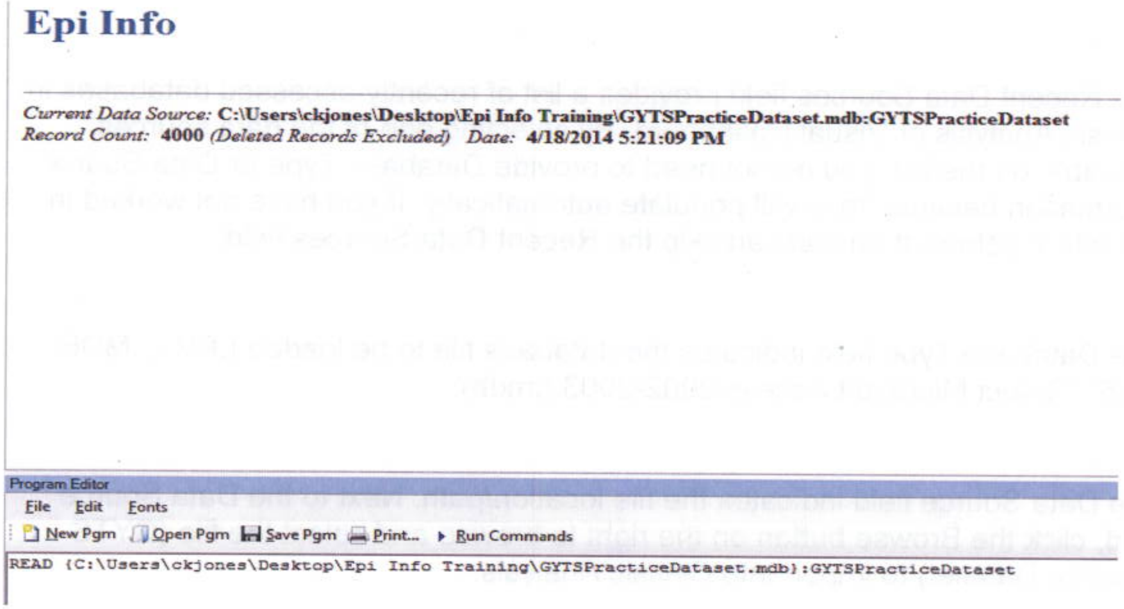
Database Type: Microsoft Access 2002-2003 (.mdb)

Data Source: C:\Users\fqa3\Desktop\Epi_Info_7\GYTSPracticeDataset.mdb

Show: ☑ Forms ☑ Tables

Data Source Explorer: GYTSPracticeDataset

In the Data Source Explorer window, click on the name of the data table (GYTS Practice Dataset) that you want to work with and click OK. The current file location, Record Count, and Date appear in the Classic Analysis Output window. The READ command appears in the Program Editor.



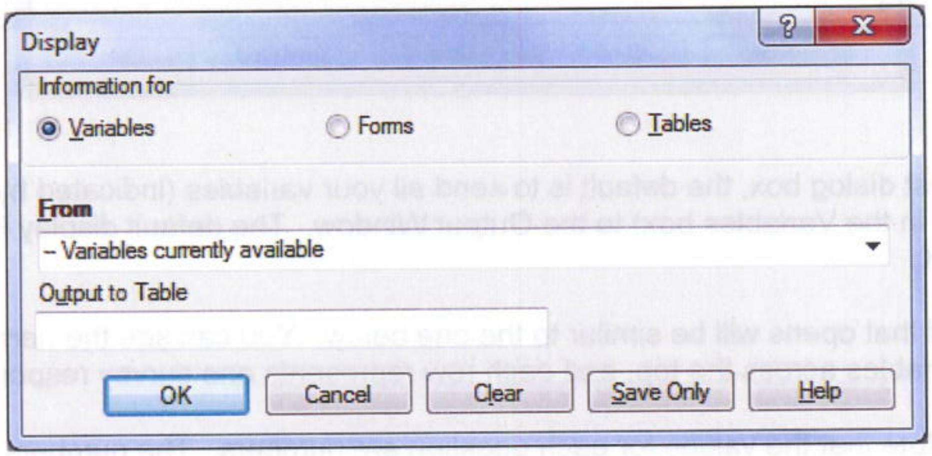
Current Data Source: **C:\Users\ckjones\Desktop\Epi Info Training\GYTSPracticeDataset.mdb:GYTSPracticeDataset**
Record Count: **4000** *(Deleted Records Excluded)* *Date:* **4/18/2014 5:21:09 PM**

```
READ {C:\Users\ckjones\Desktop\Epi Info Training\GYTSPracticeDataset.mdb}:GYTSPracticeDataset
```

Part Five: Displaying Your Variables and Data

You can view the variables in your database, also called a line listing, using the Display command. From the Classic Analysis Command window on the left, click Display under the Variables folder.

The Display dialog will then give you a chance to view the Variables. Leave the default options selected (Information for "Variables" from "Variables Currently Available") and click "OK".



Display

Information for

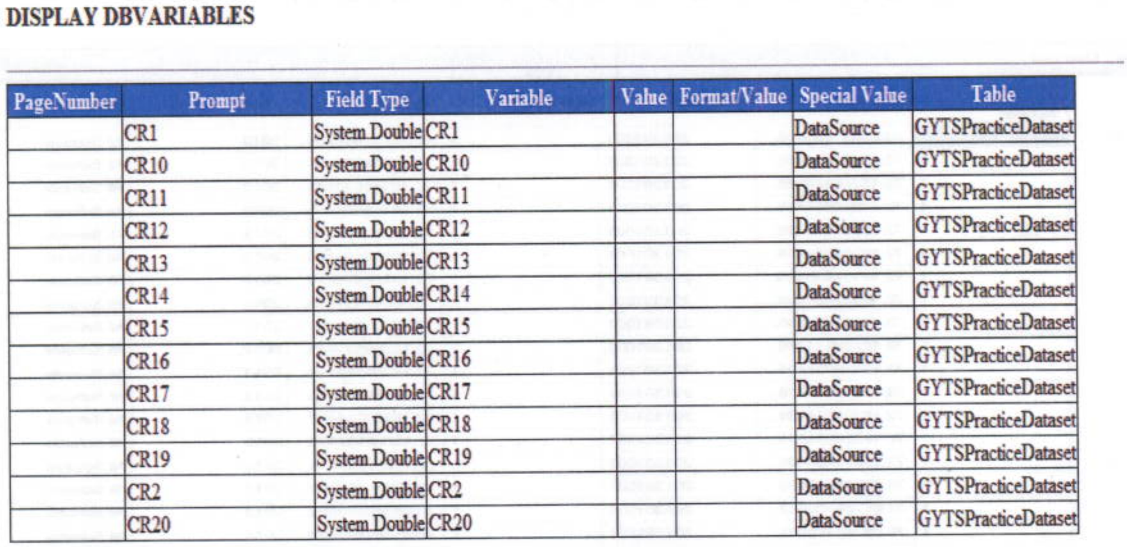
● Variables ○ Forms ○ Tables

From

-- Variables currently available

Output to Table

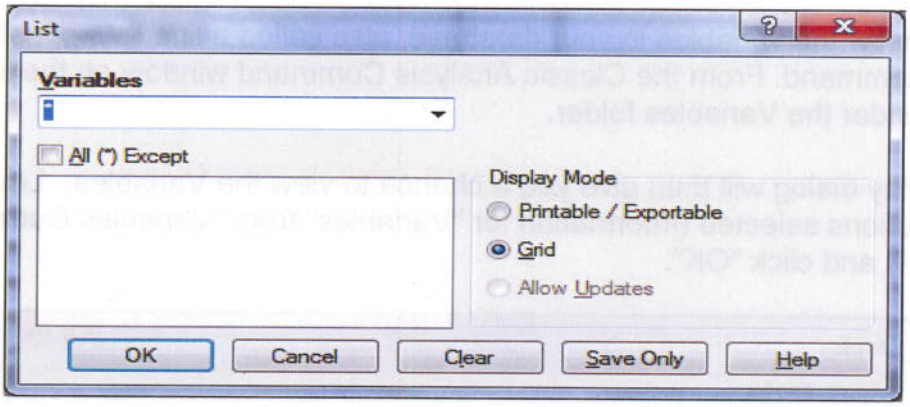
The output you should get is pictured below.



DISPLAY DBVARIABLES

PageNumber	Prompt	Field Type	Variable	Value	Format/Value	Special Value	Table
	CR1	System.Double	CR1			DataSource	GYTSPracticeDataset
	CR10	System.Double	CR10			DataSource	GYTSPracticeDataset
	CR11	System.Double	CR11			DataSource	GYTSPracticeDataset
	CR12	System.Double	CR12			DataSource	GYTSPracticeDataset
	CR13	System.Double	CR13			DataSource	GYTSPracticeDataset
	CR14	System.Double	CR14			DataSource	GYTSPracticeDataset
	CR15	System.Double	CR15			DataSource	GYTSPracticeDataset
	CR16	System.Double	CR16			DataSource	GYTSPracticeDataset
	CR17	System.Double	CR17			DataSource	GYTSPracticeDataset
	CR18	System.Double	CR18			DataSource	GYTSPracticeDataset
	CR19	System.Double	CR19			DataSource	GYTSPracticeDataset
	CR2	System.Double	CR2			DataSource	GYTSPracticeDataset
	CR20	System.Double	CR20			DataSource	GYTSPracticeDataset

You can view the data in your database using the List command. From the Classic Analysis Command window on the left, click List under the Statistics folder.

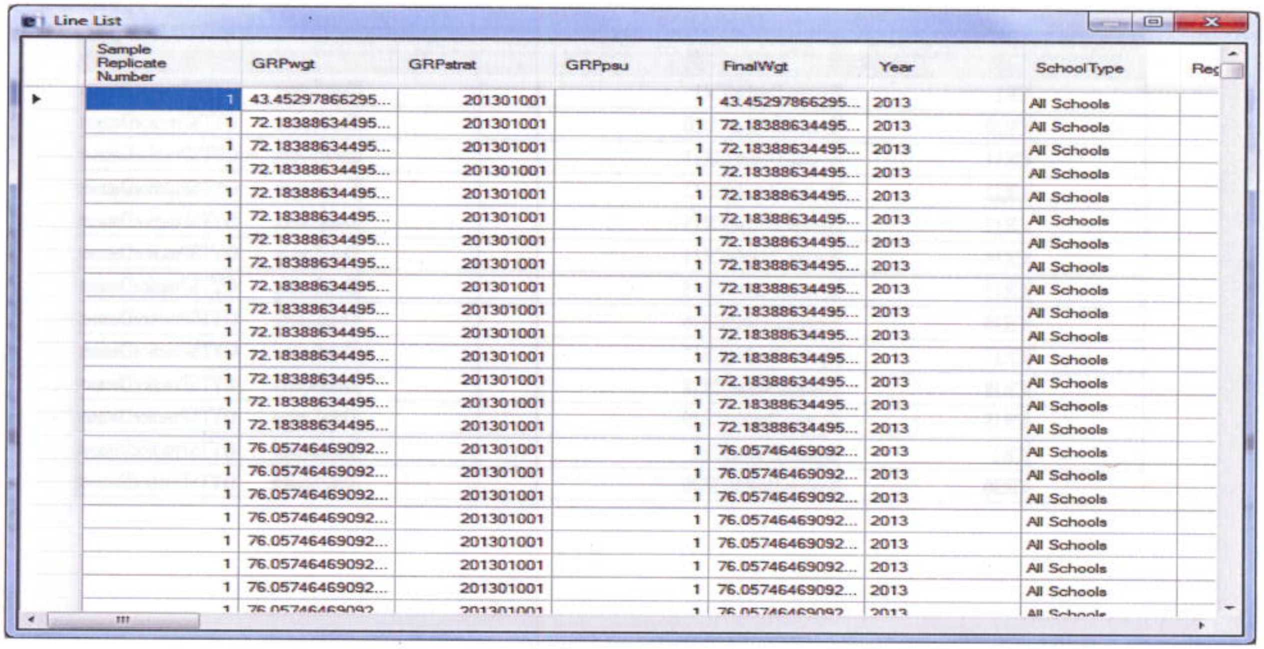


In the List dialog box, the default is to send all your variables (indicated by the asterisk in the Variables box) to the Output Window. The default display is Grid. Click OK.

The grid that opens will be similar to the one below. You can see the names of your variables across the top, and each row represents one survey respondent.

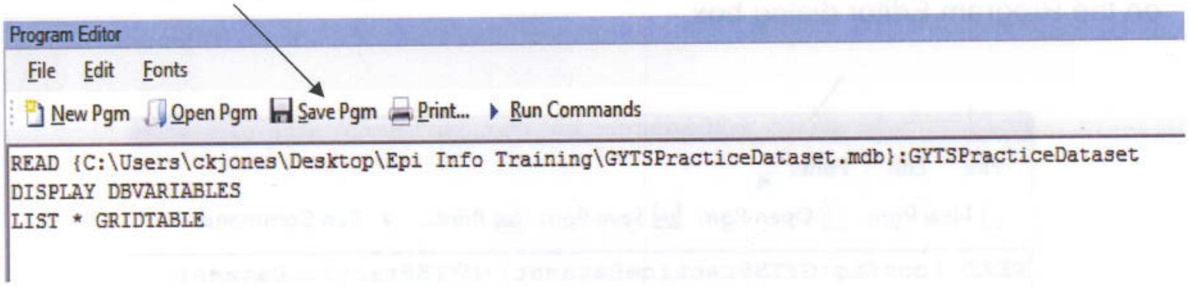
- Note that the values for each question are numbers. The numbers represent letters of the answer choices on your questionnaire:

1="A", 2="B", 3="C", 4="D", 5="E", 6="F", 7="G", 8="H"



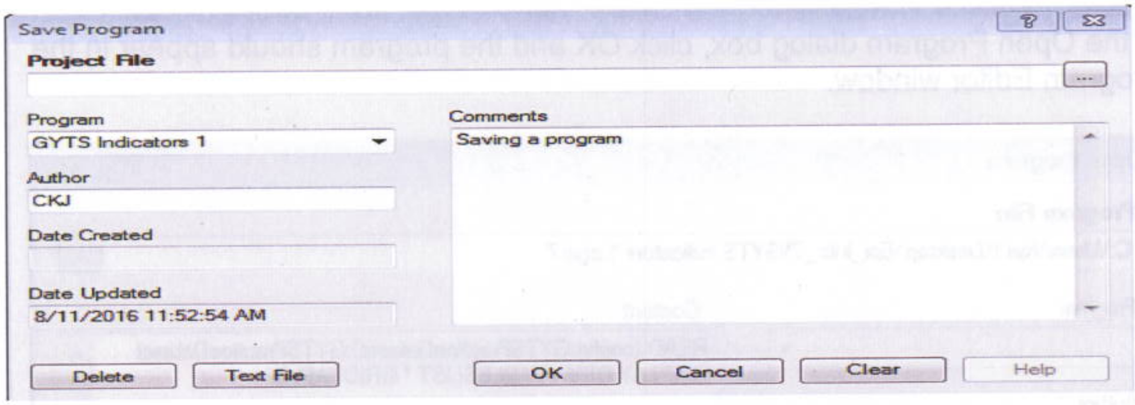
Sample Replicate Number	GRPwgt	GRPstrat	GRPpsu	FinalWgt	Year	SchoolType	Reg
1	43.45297866295...	201301001	1	43.45297866295...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	72.18388634495...	201301001	1	72.18388634495...	2013	All Schools	
1	76.05746469092...	201301001	1	76.05746469092...	2013	All Schools	
1	76.05746469092...	201301001	1	76.05746469092...	2013	All Schools	
1	76.05746469092...	201301001	1	76.05746469092...	2013	All Schools	
1	76.05746469092...	201301001	1	76.05746469092...	2013	All Schools	
1	76.05746469092...	201301001	1	76.05746469092...	2013	All Schools	
1	76.05746469092...	201301001	1	76.05746469092...	2013	All Schools	
1	76.05746469092...	201301001	1	76.05746469092...	2013	All Schools	
1	76.05746469092	201301001	1	76.05746469092	2013	All Schools	

Finally, we are going to save the commands in the Program Editor window, which altogether represent a program. The programs are stored inside your project, along with your data. Click the "Save Pgm" button on the Program Editor window.



```
READ {C:\Users\ckjones\Desktop\Epi Info Training\GYTSPracticeDataset.mdb}:GYTSPracticeDataset
DISPLAY DBVARIABLES
LIST * GRIDTABLE
```

The "Save Program" dialog box will then open. Go down to the program field, enter a program name ("GYTS Indicators 1"). Also add an author name or initials below and any comments you want on the right, including a description of the program. (The Date created and Date updated boxes are informational only, so you do not enter anything in these two areas.)



Project File

Program: GYTS Indicators 1

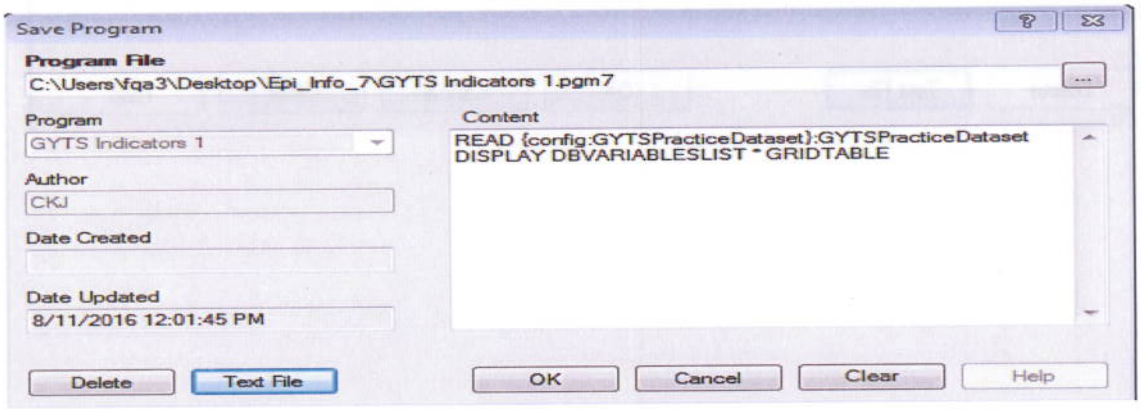
Author: CKJ

Date Created:

Date Updated: 8/11/2016 11:52:54 AM

Comments: Saving a program

Next, click Text File at the bottom and a Save As command box will come up. Choose a location or folder to save the program, type in the name of the program ("GYTS Indicators 1"), and click Save. The file path will appear in the Program File box at the top. Then click OK.



Program File: C:\Users\fqa3\Desktop\Epi_Info_7\GYTS Indicators 1.pgm7

Program: GYTS Indicators 1

Author: CKJ

Date Created:

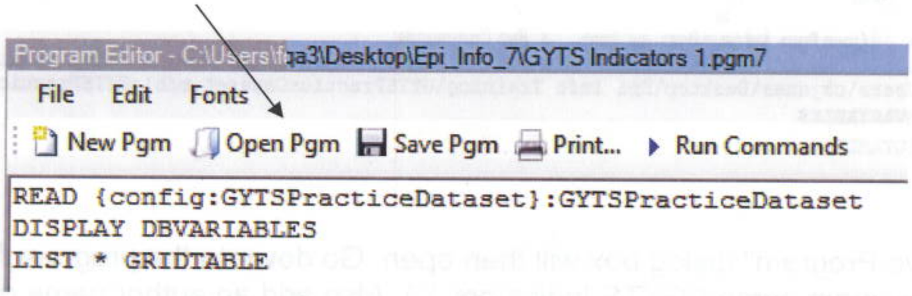
Date Updated: 8/11/2016 12:01:45 PM

Content:

```
READ {config:GYTSPracticeDataset}:GYTSPracticeDataset
DISPLAY DBVARIABLESLIST *GRIDTABLE
```

Part Six. Running a Saved Program

In Classic Analysis, open your saved program by clicking the Open Pgm button on the Program Editor dialog box.

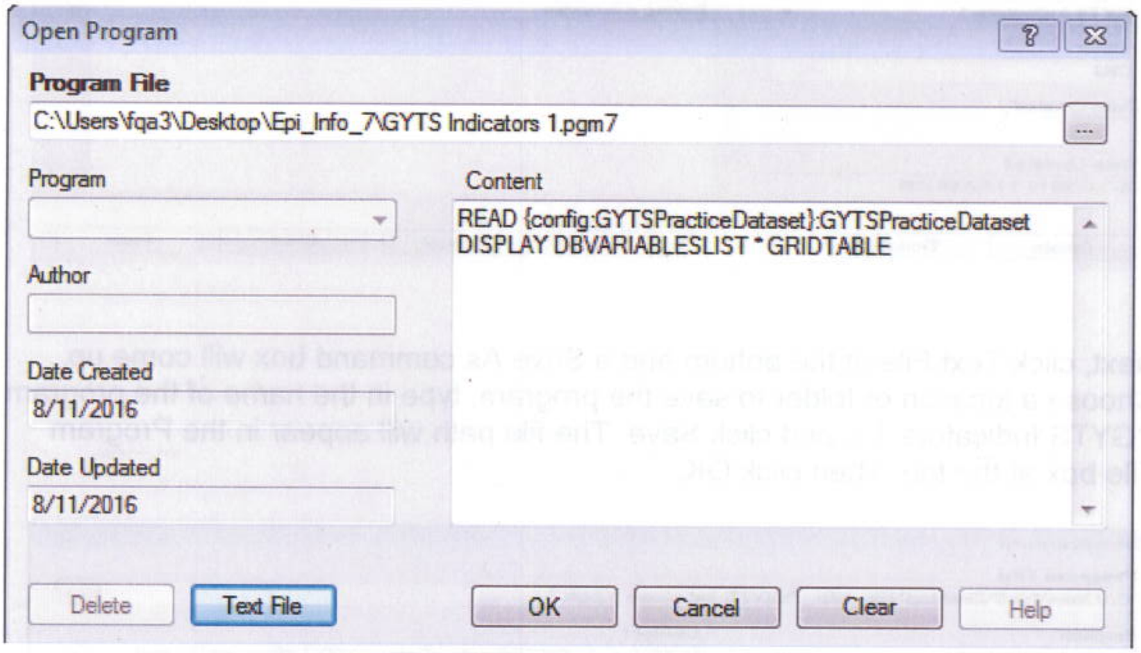


Program Editor - C:\Users\fqa3\Desktop\Epi_Info_7\GYTS Indicators 1.pgm7
File
Edit
Fonts
New Pgm
Open Pgm
Save Pgm
Print...
Run Commands

```
READ {config:GYTSPracticeDataset}:GYTSPracticeDataset
DISPLAY DBVARIABLES
LIST * GRIDTABLE
```

Click Text File at the bottom and go to the location that the program is saved. Click on the program name and then click Open.

In the Open Program dialog box, click OK and the program should appear in the Program Editor window.



Program File: C:\Users\fqa3\Desktop\Epi_Info_7\GYTS Indicators 1.pgm7

Program:

Author:

Date Created: 8/11/2016

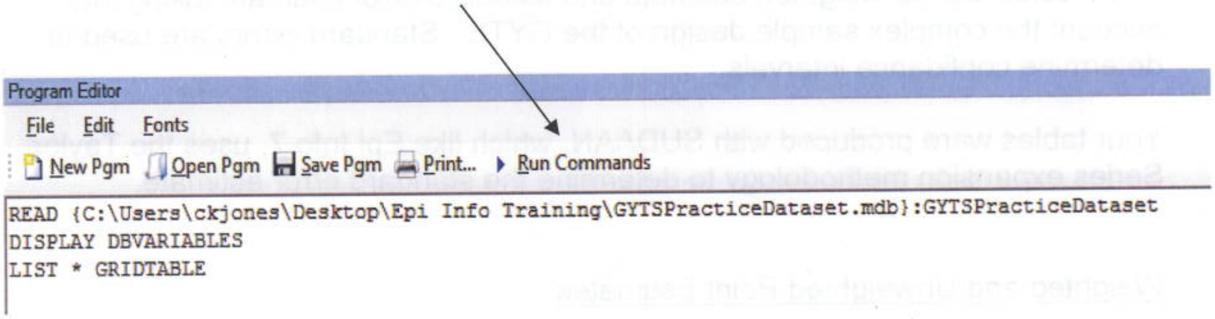
Date Updated: 8/11/2016

Content:

```
READ {config:GYTSPracticeDataset}:GYTSPracticeDataset
DISPLAY DBVARIABLESLIST * GRIDTABLE
```

Click **Run Commands** from the Program Editor window.

> *Note:* *You can run one line of command by highlighting the command with your mouse and clicking on "**Run Commands**".*



```
READ {C:\Users\ckjones\Desktop\Epi Info Training\GYTSPracticeDataset.mdb}:GYTSPracticeDataset
DISPLAY DBVARIABLES
LIST * GRIDTABLE
```

Part Seven. Weighted and Unweighted Results

The GYTS is a two-stage cluster sample. The Complex Sample functions in Epi Info 7 calculate the weighted estimate and standard error estimate taking into account the complex sample design of the GYTS. Standard errors are used to determine confidence intervals.

Your tables were produced with SUDAAN, which like Epi Info 7, uses the Taylor Series expansion methodology to determine the standard error estimate.

Weighted and Unweighted Point Estimates

WEIGHTED Point Estimates – estimates (such as a proportion or mean) produced using sampling weights to reflect the population from which the sample was drawn. Each student in the data set is assigned a sampling weight, which accounts for the following:

- Selection probability of the school
- Selection probability of the class
- Distribution of the population by grade and sex
- Non-responding schools
- Non-responding students
- Non-responding classes

UNWEIGHTED Point Estimates – these estimates only reflect the actual sample that filled out the GYTS survey. No sampling weights are used with this data.

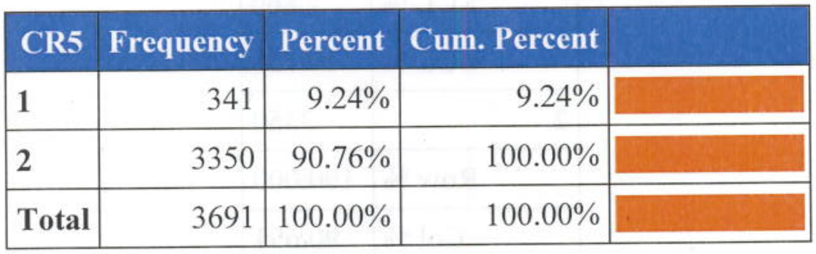
Exception: *In some situations sampling weights are not needed, such as a Census and a Simple Random Sample Survey. In these cases, each individual has the sample probability of selection.*

Running and Interpreting Unweighted Results

To obtain unweighted point estimates, click **Frequencies** under the Statistics folder. In the "Frequency of" box, choose variable CR5. Click OK.

Core Question CR5: "Have you ever tried or experimented with cigarette smoking, even one or two puffs?"

The Frequency output for core question CR5 is below.



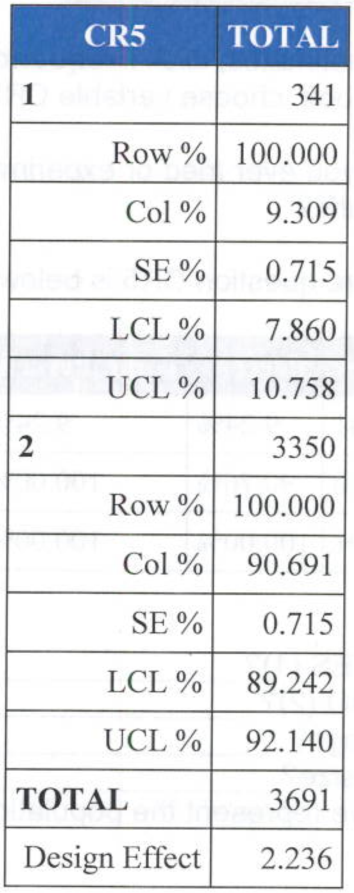
CR5	Frequency	Percent	Cum. Percent	
1	341	9.24%	9.24%	🟧
2	3350	90.76%	100.00%	🟧
Total	3691	100.00%	100.00%	🟧

How many people said YES (1)? ______________
How many people said NO (2)? ________________
What is the total percentage? _______________
What is the total sample size? _____________
Do the percentages above represent the population? ____________

Running and Interpreting Weighted Results

Click **Complex Sample Frequencies** under the Advanced Statistics folder. In the "Weight" box, choose variable Finalwgt. In the "Primary Sampling Unit" box, choose variable PSU. In the "Frequency of" box, choose variable CR5. In the "Stratify by" box, choose variable Stratum. Click OK.

The weighted Frequency output for core question CR5 is below.



CR5	TOTAL
1	341
Row %	100.000
Col %	9.309
SE %	0.715
LCL %	7.860
UCL %	10.758
2	3350
Row %	100.000
Col %	90.691
SE %	0.715
LCL %	89.242
UCL %	92.140
TOTAL	3691
Design Effect	2.236

How many people said YES (1)? ______________
What is the weighted percentage for YES (1)? ______________
What is the weighted percentage for NO (2)? ______________
What is the standard error? ______________
What is the lower confidence limit (LCL) for YES (1)? ______________
What is the upper confidence limit (UCL) for YES (1)? ______________
What is the total sample size? ______________
Do the percentages above represent the population? ______________

The weighted results show the correct percentages of the population for each answer (in this example, 9.309% YES and 90.691% NO). These are the percentages that should be used when reporting your data. Fill in the blanks below with the correct information:

__________% of the students reported that they had tried or experimented with cigarette smoking, even one or two puffs. The total sample size was __________.

Part Eight. Sampling Stages and Weights

As mentioned previously, the GYTS uses a complex sample design. To reflect the complex sample design, we have two sample design variables on your data set named **STRATUM** and **PSU** (Primary Sampling Unit).

The variable **STRATUM** usually consists of two schools. Typically, there will be two schools in each stratum, and they are paired so that both schools have similar enrollment sizes.

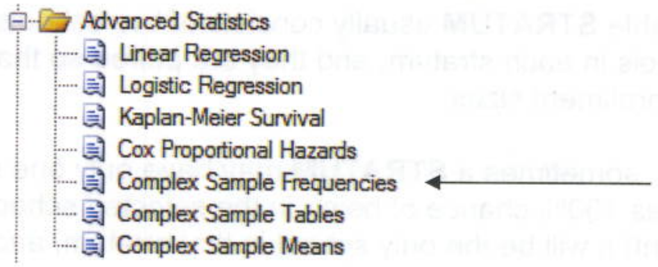
However, sometimes a **STRATUM** may have only one school. For example, if a school has 100% chance of being in the selected school list (due to large enrollment) it will be the only school in that stratum, and we call this type of school a certainty school.

In most cases, the Primary Sampling Unit represents a school. If the school is a certainty school, then the PSUs are the classes within the school.

The sampling weight variable is named **FINALWGT**.

Part Nine. Complex Sample Frequencies

In this section, we will perform complex sample analyses. To reiterate, the GYTS uses a complex sampling design and NOT a simple random sample design; hence we need to use the **Complex Sample Frequencies** or the **Complex Sample Tables** commands under the **Advanced Statistics** folder of the Classic Analysis module to obtain the correct estimates.

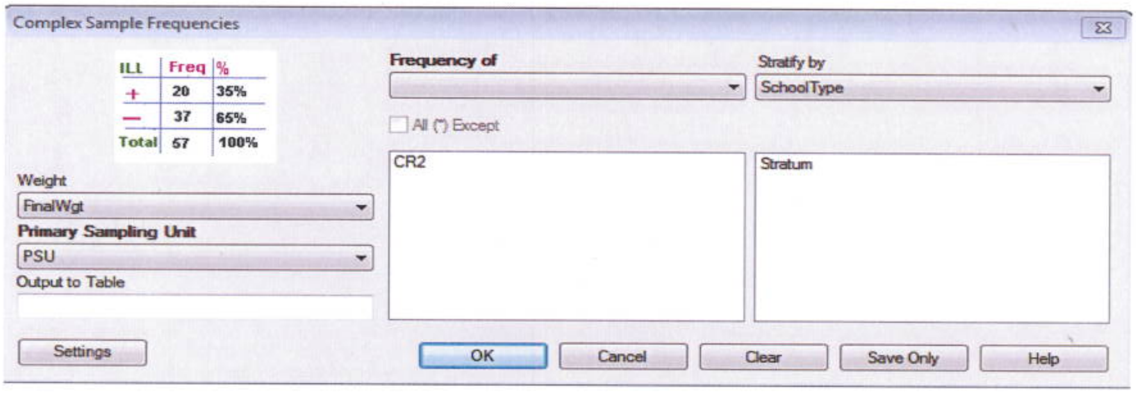


***Do not use the **Frequency** or **Tables** commands under the **Statistics** folder to obtain the estimates!

Click **Complex Sample Frequencies** under the Advanced Statistics folder, and you will see the following dialog box. We will need to fill out the following boxes:

Frequency of – identifies the variable that you would like to analyze
Weight – identifies the sampling weights variable
PSU – identifies the primary sampling unit of the survey
Stratify by – identifies the variable that represents the non-overlapping and complete groups into which the frame is divided

Using the appropriate dropdown boxes select **Finalwgt** for the "Weight" box, select **PSU** for the "PSU" box, and select **Stratum** from the "Stratify by" box. Lastly, we need to select a variable to analyze. Select **CR2** from the "Frequency Of" drop down box and then click **"OK"**.



Complex Sample Frequencies

ILL	Freq	%
+	20	35%
—	37	65%
Total	57	100%

Weight: FinalWgt

Primary Sampling Unit: PSU

Output to Table:

Frequency of:

□ All (*) Except

CR2

Stratify by: SchoolType

Stratum

Below is the output from our Complex Sample Frequencies procedure.

FREQ CR2 STRATAVAR = Stratum WEIGHTVAR = FinalWgt PSUVAR = PSU

CR2	TOTAL
1	1987
Row %	100.000
Col %	52.584
SE %	1.347
LCL %	49.855
UCL %	55.313
2	1952
Row %	100.000
Col %	47.416
SE %	1.347
LCL %	44.687
UCL %	50.145
TOTAL	3939
Design Effect	2.865

In the output above:

What is the variable analyzed? ______________________________
What is the weight variable? ______________________________
What is the PSU variable? ______________________________
What is the stratification variable? ______________________________
What is the total sample size? ______________________________
How many people answered YES? ______________________________

Now, run complex sample frequencies on the following questions from the practice dataset.

A. During the past 30 days, did you use any form of smoked tobacco products other than cigarettes (C10)?
B. Has a person working for a tobacco company ever offered you a free tobacco product (C38)?
C. How old are you (C1)?

Results for A.

Item	YES	NO	TOTAL
Weighted Percent			
Sample Size			
Standard Error			
Lower Confidence Limit			
Upper Confidence Limit			

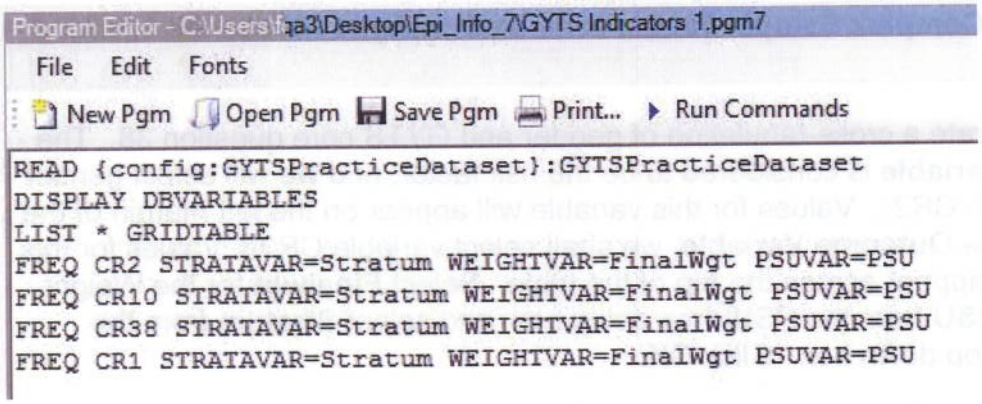
Results for B.

Item	YES	NO	TOTAL
Weighted Percent			
Sample Size			
Standard Error			
Lower Confidence Limit			
Upper Confidence Limit			

Results for C.

Item	< =11	12	13	14	15	16	17+
Weighted Percent							
Sample Size							
Standard Error							
Lower Confidence Limit							
Upper Confidence Limit							

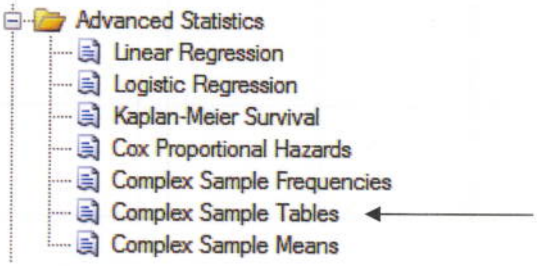
From the Program Editor window, save your program.


Program Editor - C:\Users\fqa3\Desktop\Epi_Info_7\GYTS Indicators 1.pgm7
File
Edit
Fonts
New Pgm
Open Pgm
Save Pgm
Print...
Run Commands

```
READ {config:GYTSPracticeDataset}:GYTSPracticeDataset
DISPLAY DBVARIABLES
LIST * GRIDTABLE
FREQ CR2 STRATAVAR=Stratum WEIGHTVAR=FinalWgt PSUVAR=PSU
FREQ CR10 STRATAVAR=Stratum WEIGHTVAR=FinalWgt PSUVAR=PSU
FREQ CR38 STRATAVAR=Stratum WEIGHTVAR=FinalWgt PSUVAR=PSU
FREQ CR1 STRATAVAR=Stratum WEIGHTVAR=FinalWgt PSUVAR=PSU
```

Part Ten. Complex Sample Tables

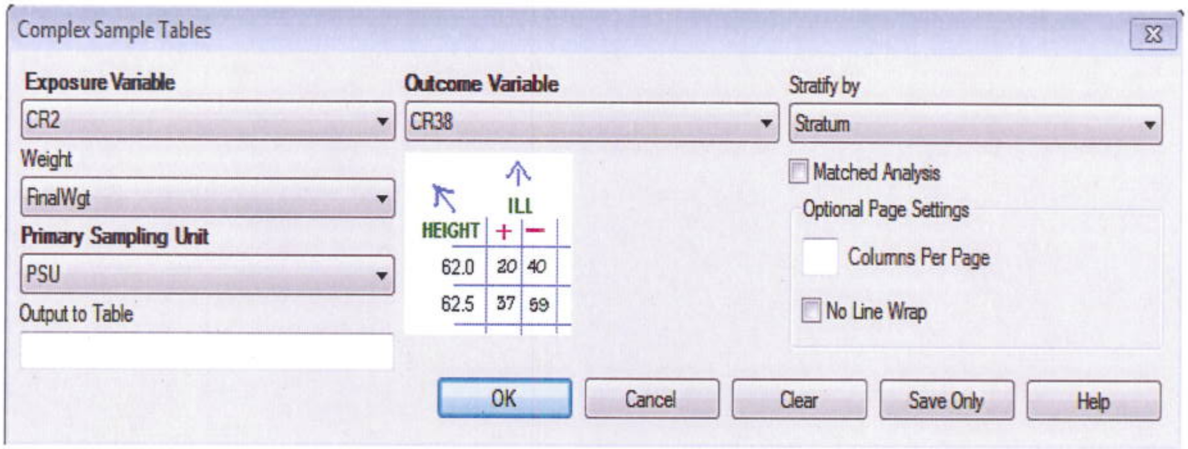
Use the **Complex Sample Tables** command from the **Advanced Statistics** folder to run cross-tabulations of variables. This command will give the correct estimates, since it takes into account the complex sample design of the survey. DO NOT use the **Tables** command under the Statistics folder since it assumes that the sample was drawn from a simple random sample.



This command is similar to the **Complex Sample Frequency** command where you will need to identify the **Weight** variable, **PSU** variable, and the **Stratify by** variable as well as the variables you would like to cross-tabulate.

Choose the **Complex Sample Tables** command, and you will see the following dialog box.

We will illustrate a cross-tabulation of gender and GYTS core question 38. The **Exposure Variable** is considered to be the risk factor, and we will select gender (core variable CR2). Values for this variable will appear on the left margin of the table. For the **Outcome Variable**, we shall select variable CR38. Values for this variable will appear across the top of the table. Select **Finalwgt** for the Weight box, select **PSU** from the PSU drop down box, and select **Stratum** from the Stratify by drop down box. Click **OK.**


Complex Sample Tables

Exposure Variable: CR2

Outcome Variable: CR38

Stratify by: Stratum

Weight: FinalWgt

Primary Sampling Unit: PSU

Output to Table:

□ Matched Analysis

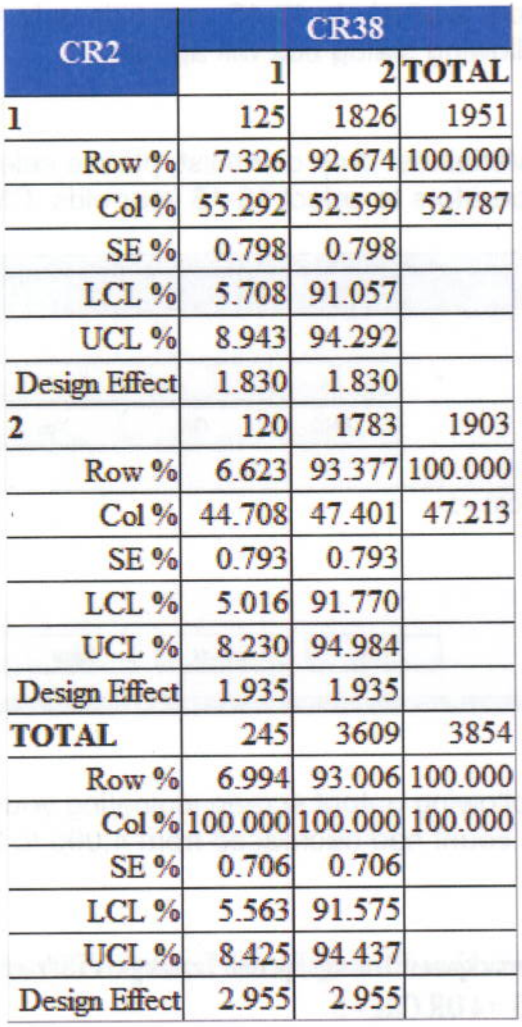
Optional Page Settings

Columns Per Page:

□ No Line Wrap

Below is the output for our cross-tabulation for CR2 by CR38.

TABLES CR2 CR38 STRATAVAR = Stratum WEIGHTVAR = FinalWgt PSUVAR = PSU



CR2	CR38		
	1	2	TOTAL
1	125	1826	1951
Row %	7.326	92.674	100.000
Col %	55.292	52.599	52.787
SE %	0.798	0.798	
LCL %	5.708	91.057	
UCL %	8.943	94.292	
Design Effect	1.830	1.830	
2	120	1783	1903
Row %	6.623	93.377	100.000
Col %	44.708	47.401	47.213
SE %	0.793	0.793	
LCL %	5.016	91.770	
UCL %	8.230	94.984	
Design Effect	1.935	1.935	
TOTAL	245	3609	3854
Row %	6.994	93.006	100.000
Col %	100.000	100.000	100.000
SE %	0.706	0.706	
LCL %	5.563	91.575	
UCL %	8.425	94.437	
Design Effect	2.955	2.955	

In the output above:

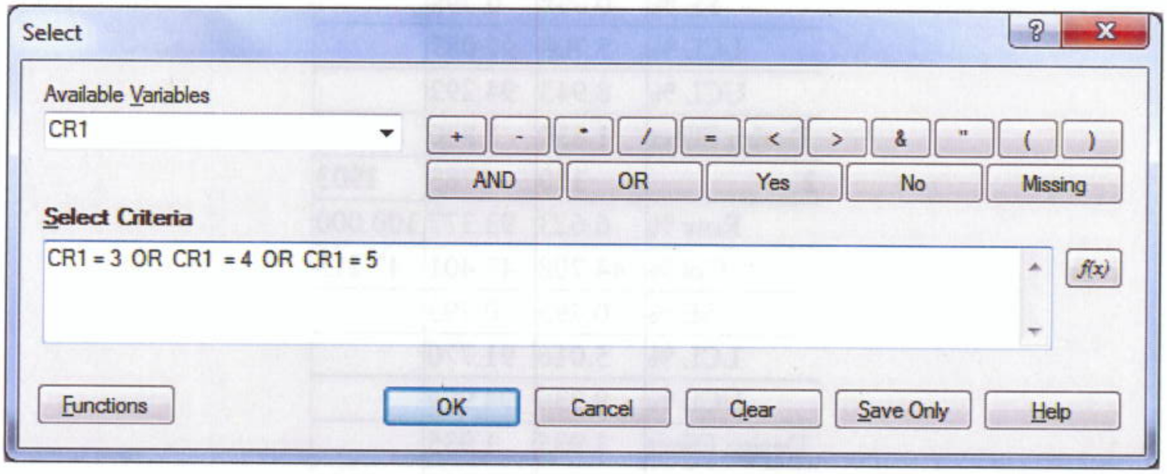
What variables are analyzed? ____________________
What is the total sample size for these two variables? ____________
How many males (CR2=1) answered YES (1) to CR38? ____________
How many total males are there? ____________________
How many males answered NO to CR38? ____________________
What is the standard error for females (CR2=2) answering YES (1)? ________
How many students answered YES to CR38? ____________________
What are the lower and upper confidence limits for females answering YES?

____

Now we would like to run cross-tabulations for respondents aged 13-15 years old. We will have to restrict the age category to ages 13-15 only.

In order to do this, we can either create a new variable or we can use the **Select** command to restrict our analysis to 13-15 year olds only. Click on the **Select** command and the following dialog box will appear.

From the "Available Variables" drop down list, please select the age variable (CR1) and use the operators to select 13-15 year olds. Click OK.



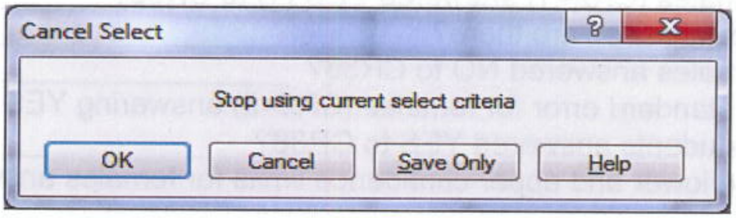
Available Variables: CR1

Select Criteria: CR1 = 3 OR CR1 =4 OR CR1 = 5

You should see the following output screen indicating your selected criteria. Note that your record count had decreased from 4,000 to 2,199.

```
Current Data Source: C:\Users\ckjones\Desktop\Epi Info Training\GYTSPracticeDataset.mdb:GYTSPracticeDataset
Selection:  CR1 = 3 OR CR1 = 4 OR CR1 = 5
Record Count:  2199 (Deleted Records Excluded)  Date:  4/29/2014 12:10:20 PM
```

Note: To remove the restriction, click Cancel Select in the Select/If folder and click OK.



Stop using current select criteria

Now perform the **Complex Sample Tables** command on CR1 (How old are you?) by CR38 and fill in the table below.

	Yes = 1	No = 2
Age 13 = 3		
%		
95% CI		
N		
Age 14 = 4		
%		
95% CI		
n		
Age 15 = 5		
%		
95% CI		
n		
Total		
%		
95% CI		
n		

Use the above procedures and **Complex Sample Tables** command to perform a cross-tabulation of gender (CR2) and core question 5 (CR5). Fill in the following table.

	Total	**Male = 1**	**Female = 2**
Yes = 1			
%			
95% CI			
n			
No = 2			
%			
95% CI			
n			
Total			
%			
95% CI			
n			

From the Program Editor window, save your program.

Part Eleven. Preferred Indicator Creation

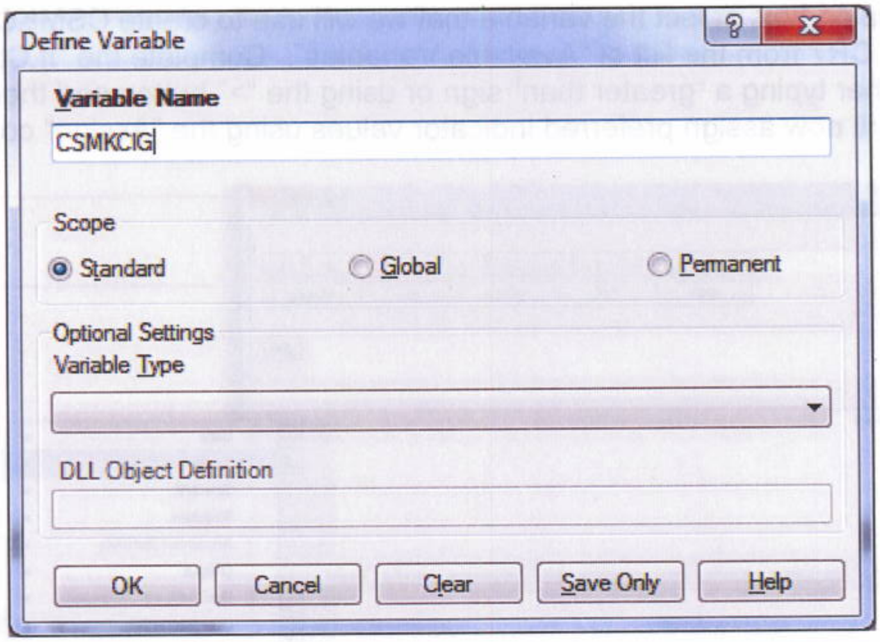
A **preferred indicator** is a variable that represents survey responses in a YES=1/NO=2 (or binary) response format. Preferred indicators are important since they are the typical results used for surveillance reporting and comparative analysis. The definitions for all of the preferred indicators are in the "GYTS Indicator Definitions" document. The definitions for all of the preferred indicators are in the "GYTS Indicator Definitions" document.

Preferred Indicator for Current Cigarette Smokers (CSMKCIG)

CSMKCIG is defined as the percentage of respondents who smoked cigarettes on 1 or more days in the past 30 days. This indicator is created using core question CR7 "During the past 30 days, on how many days did you smoke cigarettes?"

In creating the preferred indicators, we first need to use the **Define** command in the Variables folder. The Define command allows us to make new variables for our dataset. Click **Define** on the left.

You will see the dialog box shown below. Enter CSMKCIG as the new variable name and use the "Standard" scope. Click OK.



Define Variable

Variable Name: CSMKCIG

Scope: ● Standard ○ Global ○ Permanent

Optional Settings

Variable Type:

DLL Object Definition:

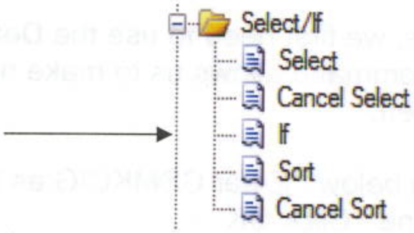
About Scope:

Scope allows the user to define the variable using one of three choices.

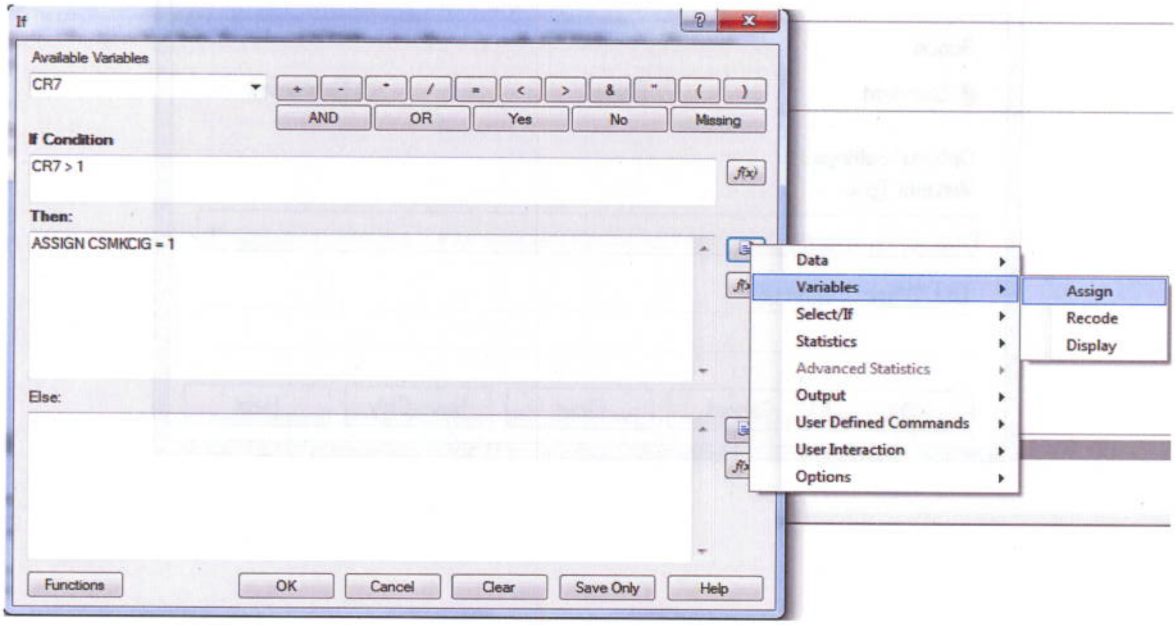
1. Standard -- These variables are kept only for the current table.
2. Global -- These variables stay alive for the whole session, even if you open another project or table.
3. Permanent -- These variables are stored in the epiinfo.ini file, and will automatically be created for every session forever.

We recommend the conservative approach of using the most limited scope possible. For the above analysis, the preferred choice is "Standard".

Now that the CSMKCIG variable has been defined, we can now assign values to it. We will designate 1 as YES, 2 as NO, and "." as missing and take three steps to create the CSMKCIG indicator. To begin, click the **If** command under the Variables folder.



In the If dialog box, select the variable that we will use to create CSMKCIG. We will select CR7 from the list of "Available Variables". Complete the "If Condition" box by either typing a "greater than" sign or using the ">" button and then typing "1". We will now assign preferred indicator values using the "Assign" command.



Available Variables: CR7

If Condition

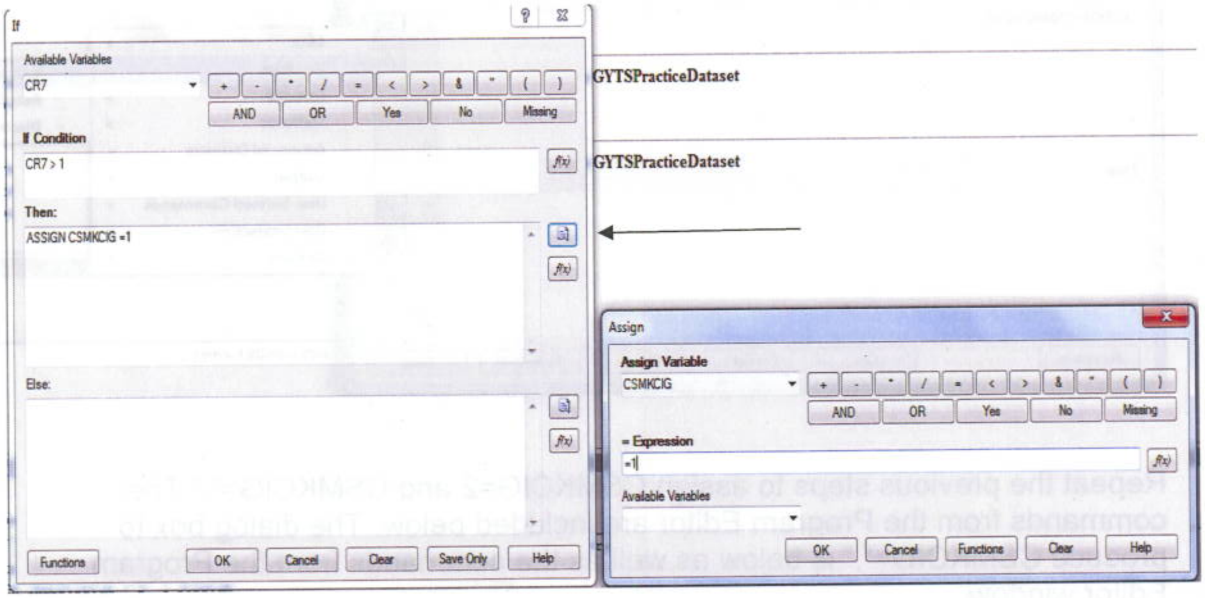
CR7 > 1

Then:

ASSIGN CSMKCIG = 1

Else:

Click inside the "Then" box. Fill in the "Then" box by either typing or choosing the Assign button on the right. The Assign dialog box will appear. Choose CSMKCIG from "Assign Variable" followed by "= 1". Click OK.



If

Available Variables: CR7

If Condition: CR7 > 1

Then: ASSIGN CSMKCIG =1

Else:

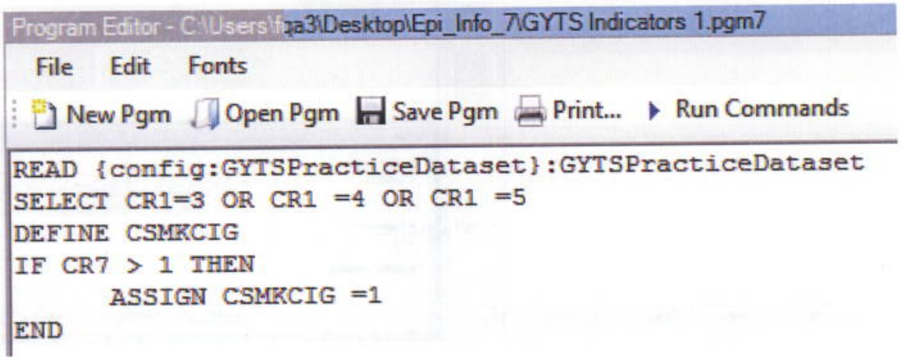
Assign

Assign Variable: CSMKCIG

= Expression: =1

Available Variables:

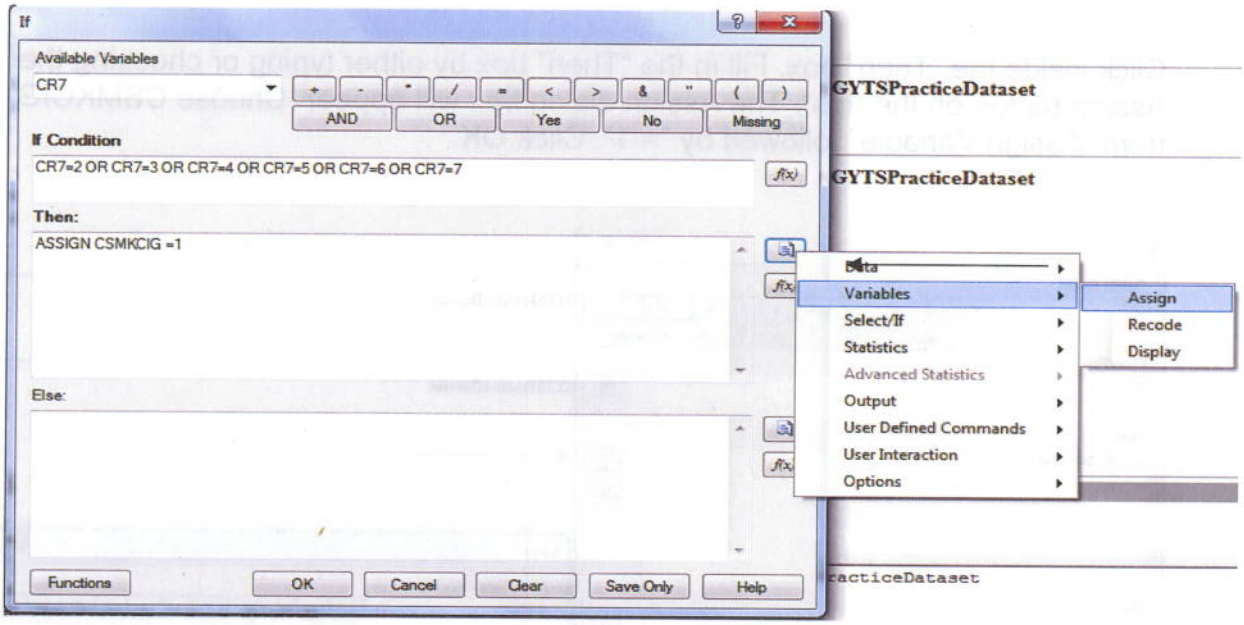
Note how the Program Editor window has changed. Some new commands have been added, including the IF/THEN commands which we just made.



Program Editor - C:\Users\fqa3\Desktop\Epi_Info_7\GYTS Indicators 1.pgm7

```
READ {config:GYTSPracticeDataset}:GYTSPracticeDataset
SELECT CR1=3 OR CR1 =4 OR CR1 =5
DEFINE CSMKCIG
IF CR7 > 1 THEN
      ASSIGN CSMKCIG =1
END
```

Note that we could have individually typed in all the individual criteria that leads to a CSMKCIG = 1 value. This alternative method also produces a correct result. In this case, the dialog box would look like the following:



If

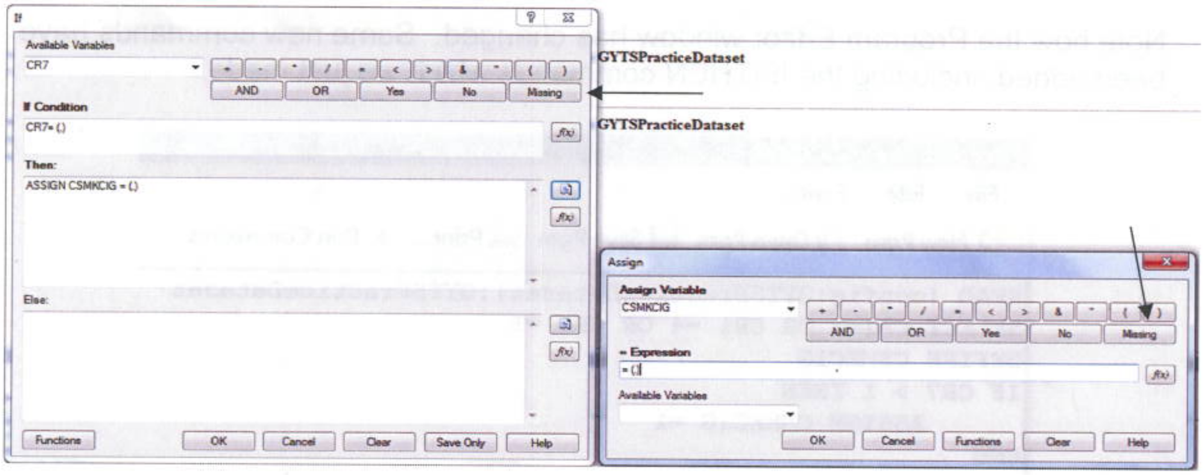
Available Variables: CR7

If Condition: CR7=2 OR CR7=3 OR CR7=4 OR CR7=5 OR CR7=6 OR CR7=7

Then: ASSIGN CSMKCIG =1

Else:

Repeat the previous steps to assign CSMKCIG=2 and CSMKCIG="." The commands from the Program Editor are included below. The dialog box to produce CSMKCIG="." is below as well as the commands from the Program Editor window.



If

Available Variables: CR7

If Condition: CR7= (.)

Then: ASSIGN CSMKCIG = (.)

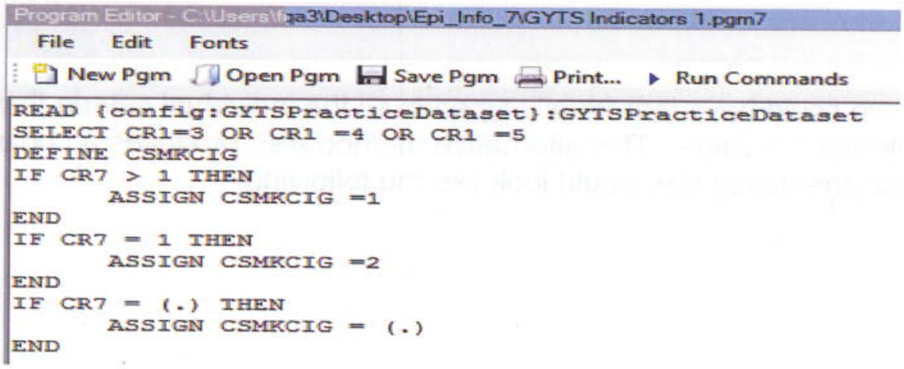
Else:

Assign

Assign Variable: CSMKCIG

= Expression: = (.)

Available Variables:



Program Editor - C:\Users\[illegible]a3\Desktop\Epi_Info_7\GYTS Indicators 1.pgm7

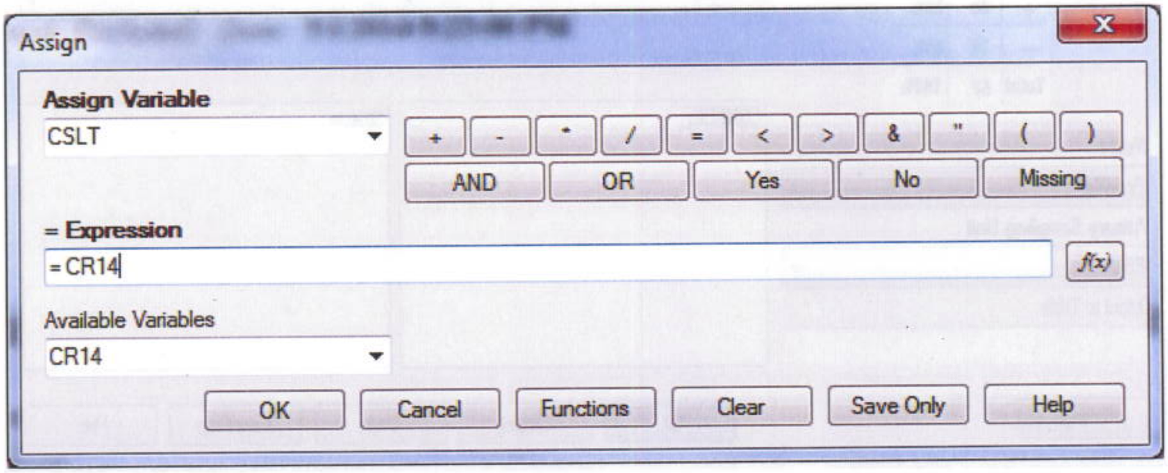
```
READ {config:GYTSPracticeDataset}:GYTSPracticeDataset
SELECT CR1=3 OR CR1 =4 OR CR1 =5
DEFINE CSMKCIG
IF CR7 > 1 THEN
        ASSIGN CSMKCIG =1
END
IF CR7 = 1 THEN
        ASSIGN CSMKCIG =2
END
IF CR7 = (.) THEN
        ASSIGN CSMKCIG = (.)
END
```

Preferred Indicator for Current Smokeless Tobacco Users (CSLT)

CSLT is defined as the percentage of respondents who used any smokeless tobacco products in the past 30 days. This indicator is created using core question CR14.

Choose the Define command on the left and enter CSLT as the new variable name. Use the "Standard" scope and click OK.

Click the Assign command under the Variables folder, and the Assign dialog box will appear. Choose CSLT from "Assign Variable" followed by "= CR14". Click OK.



Assign Variable: CSLT

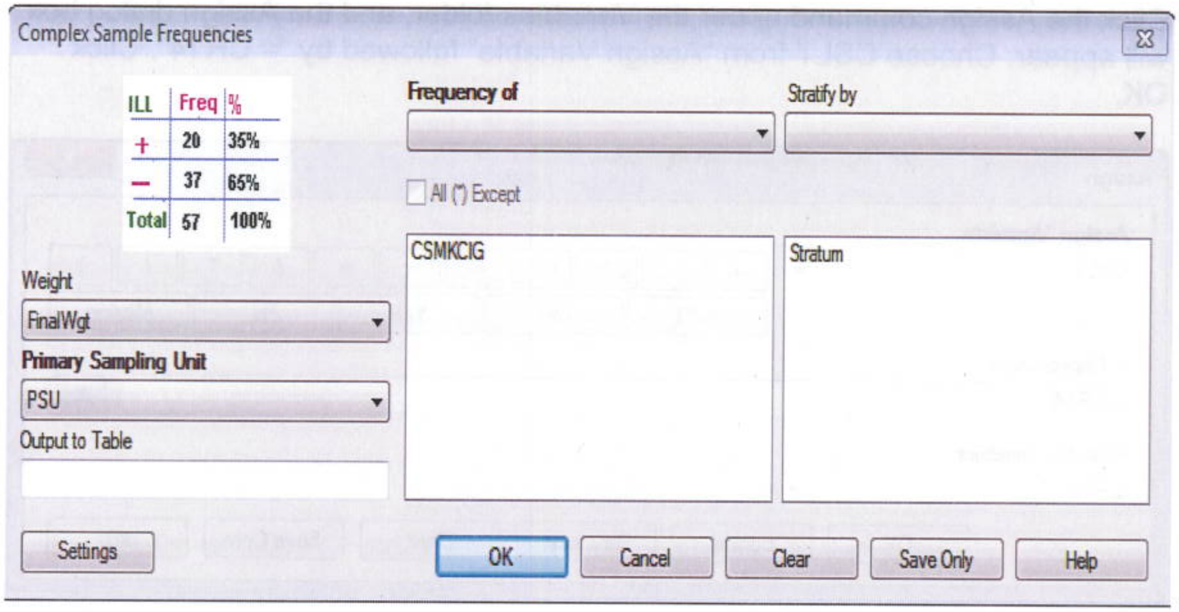
= Expression: = CR14

Available Variables: CR14

Now that you have created some of the essential preferred indicators, save your program in the Program Editor window.

Part Twelve. Preferred Indicator Frequencies

We will now use the **Complex Sample Frequencies** command that will output the correct weighted percentages, standard error, and sample sizes for the variables that you choose to run frequencies from (CSMKCIG and CSLT). Click on the **Complex Sample Frequencies** command in the Advanced Statistics folder. You will need to fill in the boxes for "Weight," "PSU," "Stratify by," and "Frequency of." Run a **Complex Sample Frequency** of the CSMKCIG variable.



Complex Sample Frequencies

Frequency of: CSMKCIG

□ All (*) Except

Stratify by: Stratum

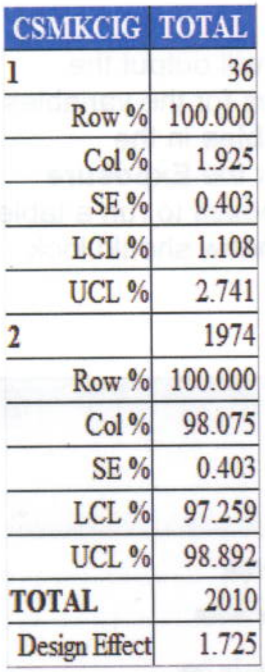
Weight: FinalWgt

Primary Sampling Unit: PSU

Output to Table:

The output from this command is below.

FREQ CSMKCIG STRATAVAR = Stratum WEIGHTVAR = FinalWgt PSUVAR = PSU



CSMKCIG	TOTAL
1	36
Row %	100.000
Col %	1.925
SE %	0.403
LCL %	1.108
UCL %	2.741
2	1974
Row %	100.000
Col %	98.075
SE %	0.403
LCL %	97.259
UCL %	98.892
TOTAL	2010
Design Effect	1.725

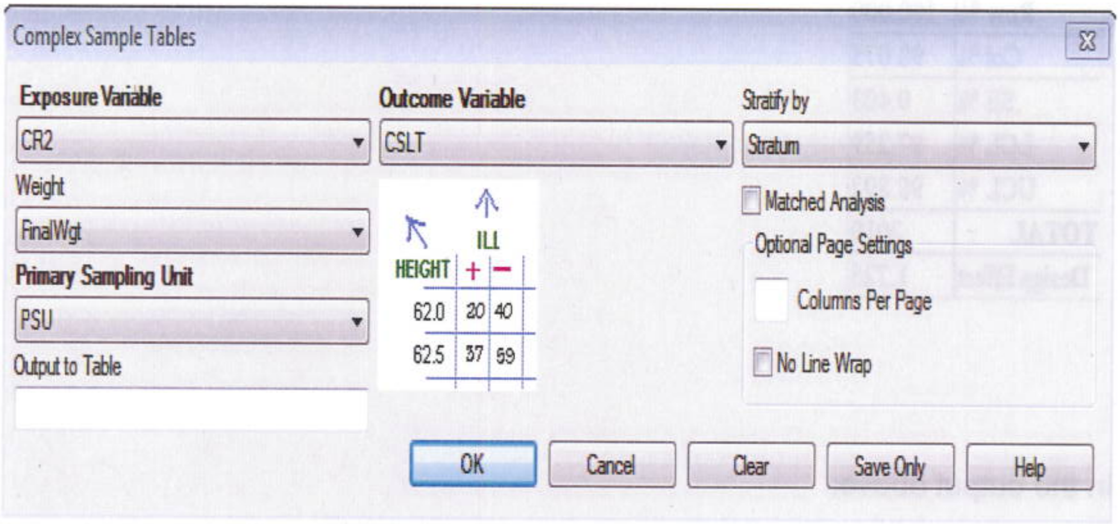
In the output above:

What is the variable analyzed? ______________________________
What is the weight variable? ________________________________
What is the PSU variable? __________________________________
What is the stratification variable? ___________________________
What is the total sample size? ______________________________
How many people are current smokers? ______________________

From the Program Editor window, save your program.

Part Thirteen. Preferred Indicator Tables

We will now use the **Complex Sample Tables** command that will output the correct weighted percentages, standard error, and sample sizes for the variables you choose for cross-tabulations. Select **Complex Sample Tables** in the Advanced Statistics folder and fill in the appropriate choices for the **Exposure Variable, Weight, PSU, Outcome Variable,** and **Stratify by** boxes to run a table on CSLT and gender (CR2). The Complex Sample Tables window should look like the following:


Complex Sample Tables

Exposure Variable: CR2

Weight: FinalWgt

Primary Sampling Unit: PSU

Output to Table:

Outcome Variable: CSLT

Stratify by: Stratum

□ Matched Analysis

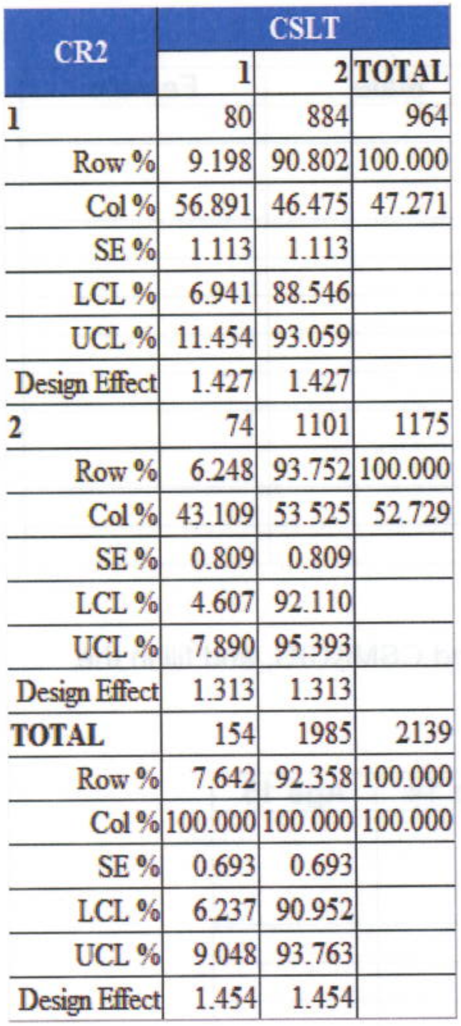
Optional Page Settings

□ Columns Per Page

□ No Line Wrap

The output from this cross-tabulation is below.

TABLES CR2 CSLT STRATAVAR = Stratum WEIGHTVAR = FinalWgt PSUVAR = PSU



CR2	CSLT		
	1	2	TOTAL
1	80	884	964
Row %	9.198	90.802	100.000
Col %	56.891	46.475	47.271
SE %	1.113	1.113	
LCL %	6.941	88.546	
UCL %	11.454	93.059	
Design Effect	1.427	1.427	
2	74	1101	1175
Row %	6.248	93.752	100.000
Col %	43.109	53.525	52.729
SE %	0.809	0.809	
LCL %	4.607	92.110	
UCL %	7.890	95.393	
Design Effect	1.313	1.313	
TOTAL	154	1985	2139
Row %	7.642	92.358	100.000
Col %	100.000	100.000	100.000
SE %	0.693	0.693	
LCL %	6.237	90.952	
UCL %	9.048	93.763	
Design Effect	1.454	1.454	

In the output above:

What are the variables analyzed? ____________________________
What is the total sample size for these two variables? ______________
In this sample, how many people are current smokers? ______________
In this sample, how many total males (CR2=1) are there? ____________
In this sample, how many males are current smokers? ______________
In this sample, how many females (CR2=2) are non-smokers? __________
What does LCL stand for? _________________________________

Using your country data, perform a Complex Sample Table on CSMKCIG and Gender (CR2) and fill in the following table. Remember you will need to restrict your analysis to ages 13 –15.

CSMKCIG – Percentage who smoked cigarettes on one or more days of the past 30 days.

	Total	**Male**	**Female**
Yes - CSMKCIG			
%			
95% CI			
n			
No - CSMKCIG			
%			
95% CI			
n			
Total			
%			
95% CI			
n			

Perform another cross tabulation of Age (CR1) and CSMKCIG, and fill in the following table.

	Total	**Age 13**	**Age 14**	**Age 15**
Yes - CSMKCIG				
%				
95% CI				
n				
No - CSMKCIG				
%				
95% CI				
n				
Total				
%				
95% CI				
n				

Perform another cross tabulation of Gender (CR2) and CSLT, and fill in the following table.

CSLT – Percentage who currently use any smokeless tobacco products.

	Total	**Male**	**Female**
Yes - CSLT			
%			
95% CI			
n			
No - CSLT			
%			
95% CI			
n			
Total			
%			
95% CI			
n			

Perform another cross tabulation of Age (CR1) and CSLT, and fill in the following tables:

CSLT – Percentage who currently use any smokeless tobacco products.

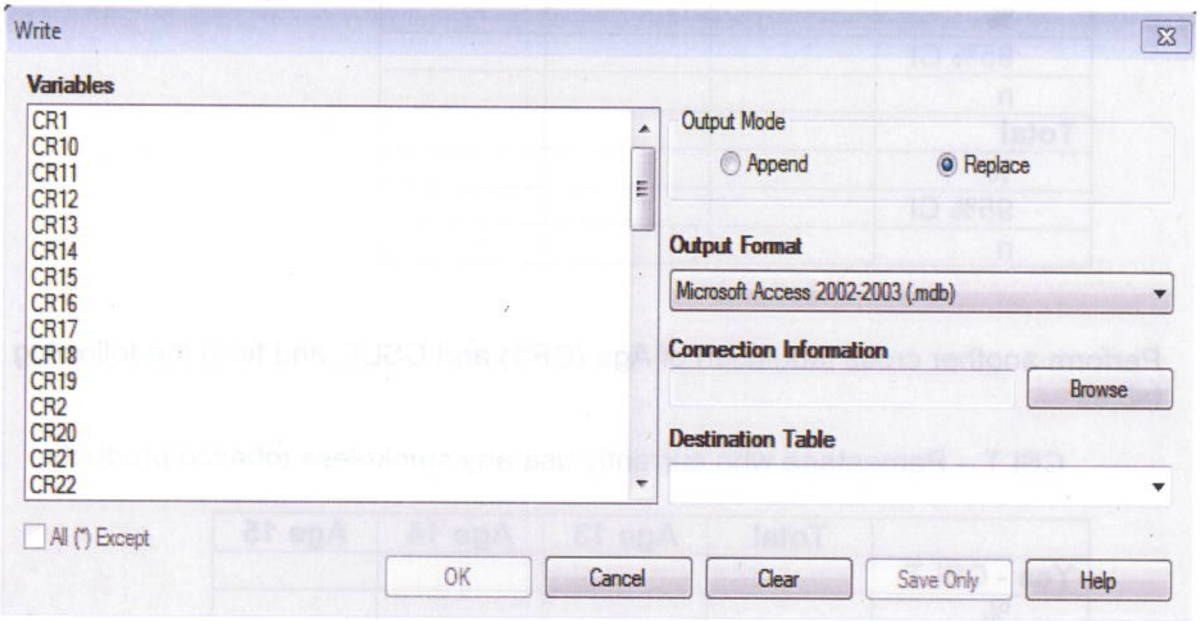
	Total	**Age 13**	**Age 14**	**Age 15**
Yes - CSLT				
%				
95% CI				
n				
No - CSLT				
%				
95% CI				
n				
Total				
%				
95% CI				
n				

Save your program using the Program Editor window.

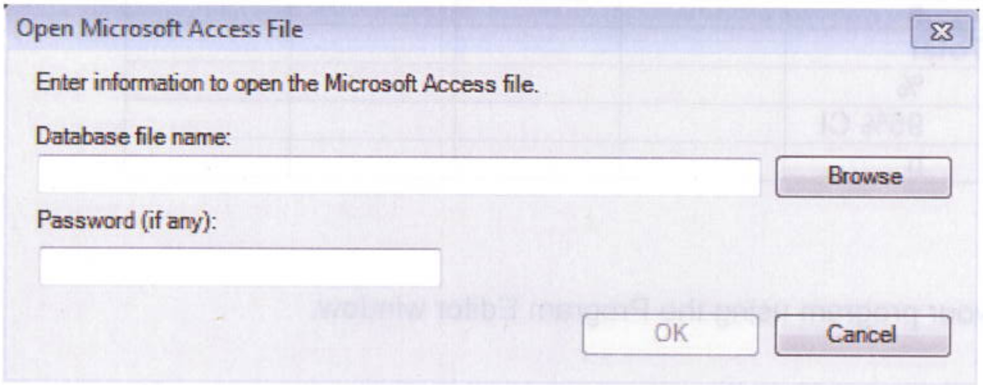
Part Fourteen. Creating a New Data Set

In this section, we will create a new data set containing the new preferred indicators. Creating a new data set or table will allow you to run analyses without having to create new variables each time.

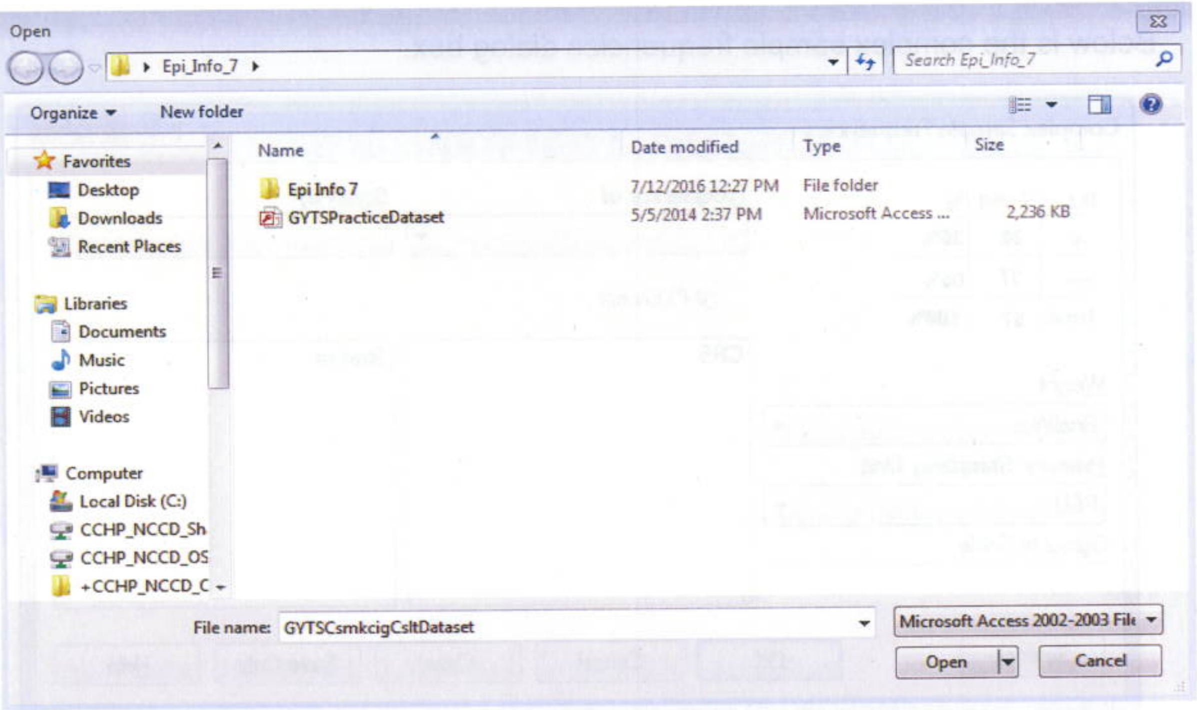
Now click on the **Write (Export**) command under the **Data** folder. The **Write Command** Window should look like the one below. Choose the **Replace** option under Output Mode.



Next to the Connection Information box, click on the Browse button and the following dialog box will appear.


Open Microsoft Access File
Enter information to open the Microsoft Access file.
Database file name:
Browse
Password (if any):
OK
Cancel

Search for the folder that you would like to save the data set in and the following window will open. Type in your file name for the data set, click Open, and then click OK.

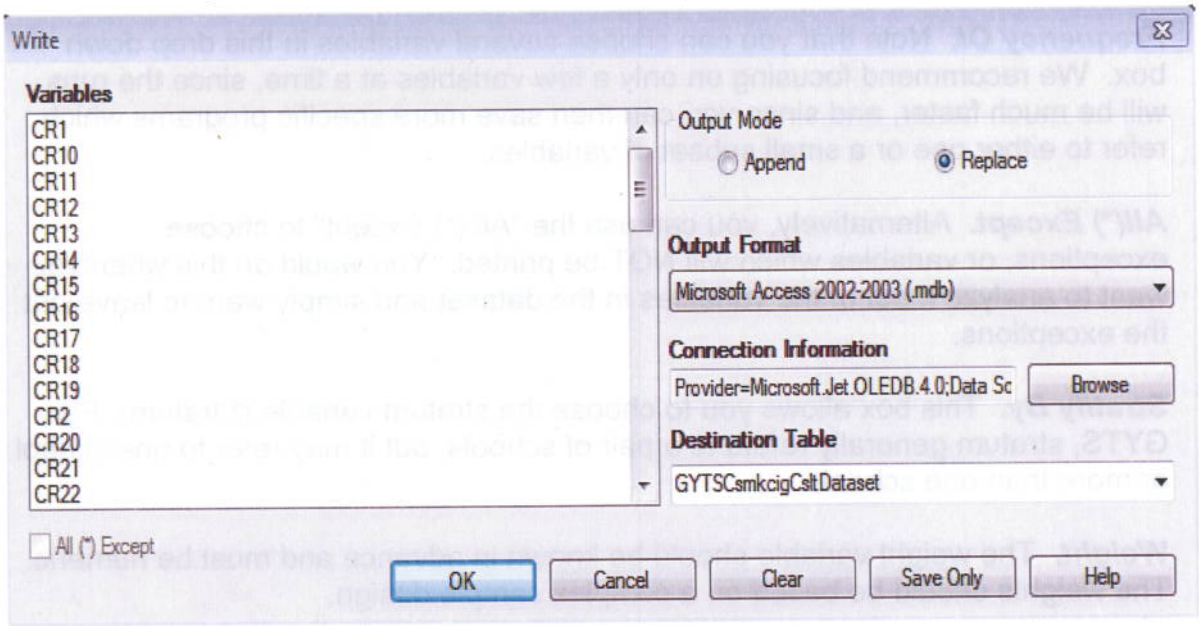


Name	Date modified	Type	Size
Epi Info 7	7/12/2016 12:27 PM	File folder	
GYTSPracticeDataset	5/5/2014 2:37 PM	Microsoft Access ...	2,236 KB

File name: GYTSCsmkcigCsltDataset

Microsoft Access 2002-2003 File

In the Destination Table field, again type the name of the data set you want to create and click OK. The Access database should save in the folder that you chose earlier.



Variables

CR1
CR10
CR11
CR12
CR13
CR14
CR15
CR16
CR17
CR18
CR19
CR2
CR20
CR21
CR22

Output Mode

○ Append ● Replace

Output Format

Microsoft Access 2002-2003 (.mdb)

Connection Information

Provider=Microsoft.Jet.OLEDB.4.0;Data Sc

Destination Table

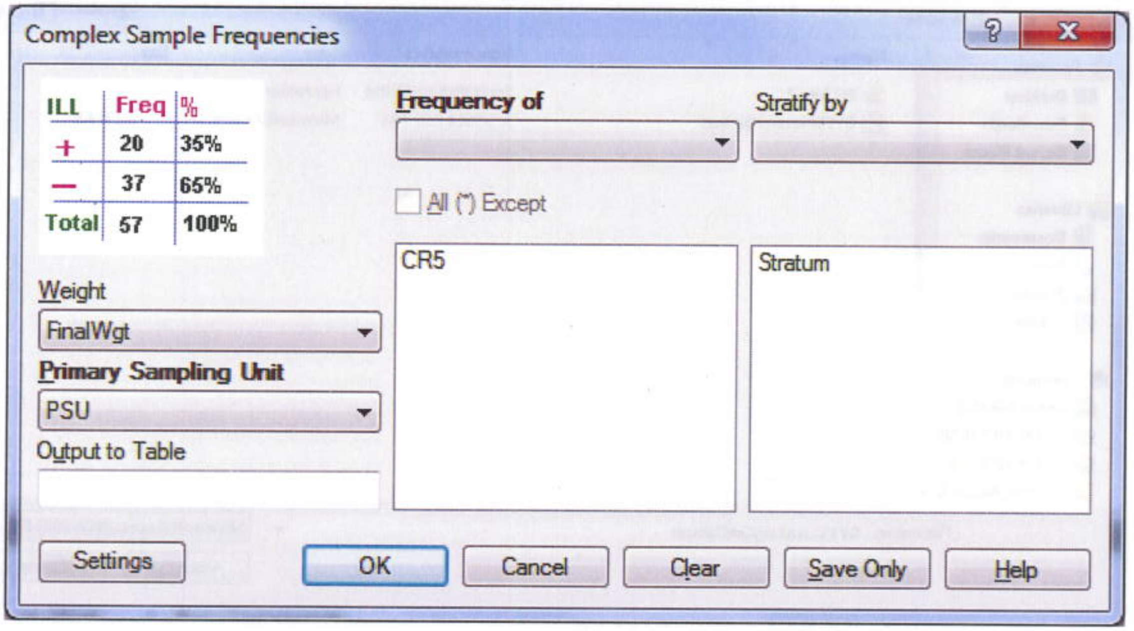
GYTSCsmkcigCsltDataset

□ All (*) Except

Appendix One. Complex Sample Frequencies

This section will discuss the features of complex sample frequencies.

Below is the complex sample frequencies dialog box.



Complex Sample Frequencies

ILL	Freq	%
+	20	35%
—	37	65%
Total	57	100%

Weight: FinalWgt

Primary Sampling Unit: PSU

Output to Table

Frequency of: CR5

□ All (*) Except

Stratify by: Stratum

"ILL" Sample Frequency Table. The upper left contains an example of the type of report which will be produced. The "ILL" refers the variable of interest and is the type of variable which would be chosen in the "Frequency Of" drop down box.

Frequency Of. Note that you can choose several variables in this drop down box. We recommend focusing on only a few variables at a time, since the runs will be much faster, and since you can then save more specific programs which refer to either one or a small subset of variables.

All(*) Except. Alternatively, you can use the "All (*) Except" to choose exceptions, or variables which will NOT be printed. You would do this when you want to analyze most of the variables in the dataset and simply want to leave out the exceptions.

Stratify By. This box allows you to choose the stratum variable (Stratum). For GYTS, stratum generally refers to a pair of schools, but it may refer to one school or more than one school.

Weight. The weight variable should be known in advance and must be numeric. The weights should be based on a complex sample design.

PSU. The acronym PSU stands for Primary Sampling Unit and refers to the smallest unit of the design. In GYTS, PSU generally refers to an individual school.

Output to Table. The program has the option to output your results to a new table. You would do this when your output is fairly complex, or you had some intention to do even further analysis on this new resulting table. You put the table name here, and then you can use the Read (Import) function (choosing "All", since this new output table is a regular table) to read in this new table.

Clear. Clears all the input boxes.

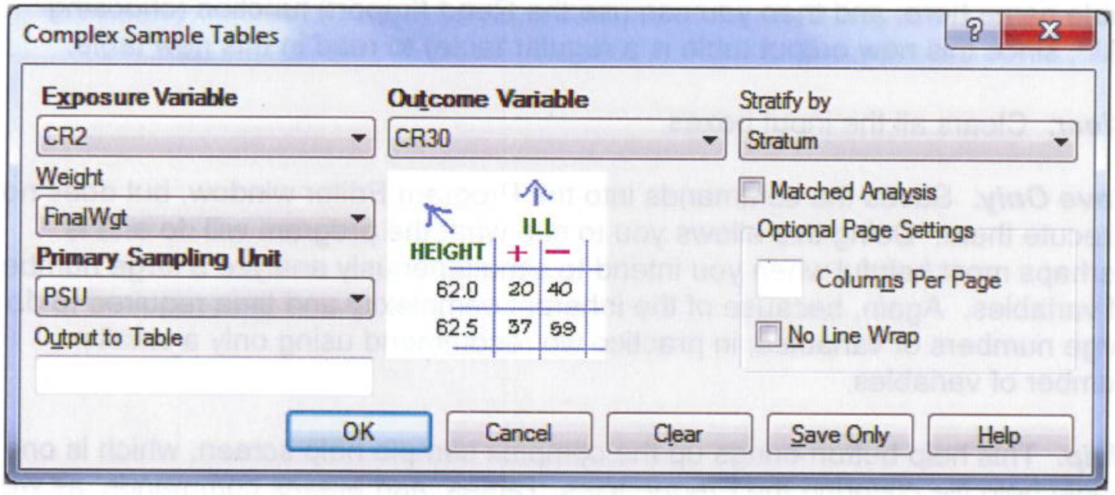
Save Only. Saves the commands into the Program Editor window, but does not execute them. Doing this allows you to see what the program will do and is perhaps most helpful when you intend to simultaneously analyze a large number of variables. Again, because of the inherent complexity and time required to do large numbers of variables, in practice we recommend using only a smaller number of variables.

Help. This help button brings up the complex sample help screen, which is one single help file covering the Frequencies, Tables, and Means commands, as well as the science of complex samples.

Appendix Two. Complex Sample Tables

This section will discuss the features of complex sample tables.

Below is the complex sample tables dialog box.


Complex Sample Tables

Exposure Variable: CR2

Outcome Variable: CR30

Stratify by: Stratum

Weight: FinalWgt

Primary Sampling Unit: PSU

Output to Table:

□ Matched Analysis

Optional Page Settings

Columns Per Page:

□ No Line Wrap

HEIGHT	+	−
62.0	20	40
62.5	37	59

"HEIGHT by ILL" Sample Table. The Height by Ill sample table is an example of what goes in the "Exposure variable" and "Outcome variable" selection boxes. The "Height" variable is the categorical independent variable, and the "Ill" variable is the categorical dependent variable. The example assumes that there are only a relative few number of different height values possible, and therefore categorical analysis is the correct method for interpreting Height. Mathematically, it does not depend which variable is which, and the output will then be based on your preference, with the "exposure variable" being displayed vertically and the "outcome variable" being displayed horizontally. Thus, you can simply choose the variable with fewer choices as the horizontal (or "outcome") variable.

Exposure Variable. This drop down box allows for one selection, and the results will display vertically.

Outcome Variable. This drop-down box allows for one selection, and the results will display horizontally. On the output, the standard error percentage will refer to the row percentage.

Weight. The weight should be proper for a complex sample design.

PSU. The acronym PSU stands for Primary Sampling Unit and refers to the smallest unit of the design. In GYTS, PSU generally refers to an individual school.

Stratify By. This box allows you to choose the stratum variable (Stratum). For GYTS, stratum generally refers to a pair of schools, but it may refer to one school or more than one school.

Output to Table. The program has the option to output your results to a new table. You would do this when your output is fairly complex, or you had some intention to do even further analysis on this new resulting table. You put the table name here, and then you can use the Read (Import) function (choosing "All", since this new output table is a regular table) to read in this new table.

Clear. Clears all the input boxes.

Save Only. Saves the commands into the Program Editor window, but does not execute them. Doing this allows you to see what the program will do, and is perhaps most helpful when you intend to simultaneously analyze a large number of variables. Again, because of the inherent complexity and time required to do large numbers of variables, in practice we recommend using only a smaller number of variables.

Help. This help button brings up the complex sample help screen, which is one single help file covering the Frequencies, Tables, and Means commands, as well as the science of complex samples.

Appendix Three. Design Effect

The Epi Info 7 manual describes the Design Effect (DEFF) as follows:

> "A useful design-related measure for surveys is the so-called "design effect" (Kish, 1965), the ratio of the variance of the estimate (under the actual design used to produce the estimate) to the variance of the estimate assuming the same data to have come from a simple random sample (this definition may differ from that used by other computer programs for complex survey analysis). The design effect reflects the estimated variance of the survey data relative to that of a simple random sample. Design effect for multi-stage cluster samples will usually exceed 1, sometimes substantially, while for stratified simple random samples and other list samples design effect will be near or slightly less than 1.
>
> Generally stratification tends to reduce the design effect and cluster sampling to increase it. Widely variable sample weights tend to increase design effect."

The statistical definition of the Design Effect (DEFF) is the following ratio:

$$DEFF = \frac{\hat{V}_d\left(\hat{\Theta}\right)}{\hat{V}_s\left(\hat{\Theta}\right)}$$

Where $\hat{V}_d$ = variance estimator based on study design

$\hat{V}_s$ = variance estimate of a simple random sample of equal size

Items to Note about the Design Effect:

- The Design Effect is useful for calculating the samples size for future surveys in which you are using the same sampling frame.
- The Design Effect is usually different for every question on a survey.
- Because the GYTS is a two stage-cluster sample, the DEFF is almost always greater than 1.
- Very large values of DEFF (greater than 10) may indicate a problem with the survey.

Glossary of Terms

Cluster sampling. A type of sampling method in which the sampling unit is at some point a group rather than an individual.

Confidence Interval (95%). A 95% Confidence Interval (CI) for a population parameter is an interval constructed from a sample so that, loosely, there is a 95% chance that the interval contains the parameter. Strictly, with repeated sampling, 95% of intervals so constructed would include the population parameter. A confidence interval provides an estimate of how accurately the population parameter has been estimated.

Design effect. The ratio of the observed standard error to the expected standard error if simple random sampling is used. See Appendix Three for more information on design effect.

Frequency Tables. Tables which contain the percentage distribution and the associated 95% confidence intervals for each question from a country's GYTS questionnaire.

Measurement. The process of obtaining the qualitative or quantitative values needed to meet research objectives.

Nonresponse error. The inability to obtain data for all questionnaire items from persons in the sample population.

p-Value. The probability of getting the observed result or one more extreme assuming that the null hypothesis is true.

Percentage. The number (of cases or people) per 100. For example, one out of 100 equals 1%; 50 out of 100 equals 50%, or half of the total number; 25% is a quarter of the total, or one in four; 75% is three out of every four in a group or 75 out of 100.

Post-stratification. The method used to adjust the distribution of the sample data so that it reflects the total population of the sampled area. The post-stratification factor is calculated by computing the ratio of the gender and grade distribution of the school enrollment population divided by that of the sample. This factor is then multiplied by the raw weight to compute an adjusted, final weight variable. The weighting adjusts not only for variation in selection and sampling probability but also for gender and grade so that projections can be made from the sample to the general population. Weighting of the sample also adjusts for nonresponse and non-coverage (failure of some schools and/or students elements to be included in the sampling frame).

Prevalence. The number of existing cases of a disease. This number includes those who have it and those newly diagnosed with it.

Probability sample. A sample in which each member of the population has a known, nonzero probability of selection.

Preferred Tables. Prevalence data indicators for GYTS which show the prevalence (percentage) for a single created variable or a combination of variables in the GYTS questionnaire.

Primary Sampling Unit (PSU). The sampling units that are selected in the first stage of a typical two-stage survey. In the GYTS, PSUs are usually a school or a pair of schools. For certainty schools, one PSU represents each class within that certainty school.

Raw data. Actual responses received from survey respondents.

Reliability. A measure of the extent to which observations of a study are repeatable or produce the same answers. Measurement unreliability may be inherent in the instrument itself (e.g., the wording of a question) or come from differences in procedure (e.g., the interviewer's tone of voice when asking the question). A question is reliable if it produces consistent responses.

Sample. A small group selected to represent a larger population.

Significance Level. An arbitrary level below which a result is deemed 'significant', that is, unlikely to have arisen by chance and as a consequence of which the null hypothesis is rejected. It is usually set at 0.05 or 5%.

Simple Random Sample. A design in which every member of the surveillance population has an equal chance of being selected to participate in the survey.

Standard Deviation (SD). A measure of variability in a population.

Standard Error (SE). A measure of the precision of a sample estimate. It is always associated with a parameter estimate such as a mean, regression coefficient or odds ratio. Note that it depends on the sample size.

Sampling fraction. The number of elements selected in a stratum divided by the number of elements considered for selection in some stage of sampling.

Stratum. A subset of sampling units treated as a single group in the selection of a sample. A typical stratum in GYTS consists of a pair of schools.

Stratum ID. A numeric code assigned to each stratum to differentiate it from another stratum. For the GYTS data, the Stratum ID is the variable STRATUM.

Unweighted. A point estimate (percentage) that was NOT calculated using the sampling weights. Unweighted estimates should NOT be used when reporting point estimates (percentages).

Weighted. An estimate (percentage) that was created using the sampling weights.

Common Abbreviations and Acronyms

CRXX	A variable name indicating the Core GYTS questionnaire number
	Examples:
	CR1 = GYTS Core Question number 1
	CR5 = GYTS Core Question number 5

FINALWGT	Final sampling weight variable on your GYTS data set
LCL %	Lower Confidence Limit. Lower bound of 95% Confidence Interval
n	Sample Size
OSH	Office on Smoking and Health at the CDC
PSU	Primary Sampling Unit variable on your GYTS data set
SE	Standard Error
STRATUM	Stratum ID variable on your GYTS data set
95% CI	95% confidence Interval
UCL %	Upper Confidence Limit. Upper bound of 95% Confidence Interval