**A Survey of Asthma in Miami-Dade Schoolchildren**

**Introduction---Sources of Data and Tools for**

**Analysis and Data Capture**

Public Health Data exist in many forms. In many cases data have been captured by others and stored in files, and the challenge is to clean up and perhaps reformat the files, and then analyze the data. This is true of data that you might download from a surveillance website, or a large national survey, or small surveys in your community.

**Epi Info** has several components (to turn a questionnaire into a data-entry screen, enter data, etc.)

* To analyze and produce statistics from data **already** in database files (whether they were created in **Epi Info** or in some other software [Excel, whatever)) we will use the program called **ANALYSIS** in **Epi Info**. We are going to cover this process in the **FIRST** session.

---------------------------------------------------------------------------------------------------------------------

* In other situations, the data have not yet been captured in a computer program. They exist in live population, “on the hoof”, patient medical or laboratory records, and must be captured and organized for analysis. You might obtain data by:
  + *Abstracting hospital records*
  + *Having participants or observers fill out questionnaires*
  + *Conducting laboratory studies, and recording the results*
  + *Many other ways!*
* This process, which might be called data capture or collection and data-entry or digitization, will be covered in the **SECOND** session of the course.

Another free program is called **OpenEpi**

* It is useful for statistical analysis when the data are in summary form (as might be the case when you) want to:
  + Check calculations in a paper or manuscript you are reviewing
  + Check calculations in a published paper or manuscript, or a presentation
  + See if differences in proportions of different groups in your community with some health condition are statistically significant
  + See if factors in populations are protective or increase risk for some health condition
* We will cover these uses of OpenEpi in the **FIRST** session of the course.
* OpenEpi has many other uses, such as sample-size calculation and random number generation, which will be covered in the **SECOND** session of the course.

**Where are we, and where are we going?**

* In this session of the course, we will start with datasets that have already been captured as Excel spreadsheet files, so you don’t have to worry about data entry. One of them is about YOU! It’s the answers that many of you provided to our pre-evaluation!
* We will import these files into Epi Info using the IMPORT command in the ANALYSIS program of Epi Info, and answer some questions about your backgrounds. You are REALLY diverse!
* From the other dataset, we will summarize aspects of asthma in schoolchildren in Miami. Please, note that the Miami asthma in schoolchildren data\* were generously provided by our collaborators in a needs assessment—the data have been scrambled and altered in random ways to protect the source of the data. They are fine for learning Epi Info but NOT for your research, presentation, publication, or even believing them! They should be regarded as FAKE!!
* The Asthma in Schoolchildren exercise assumes that you have:
  + Installed Epi Info version 3.5.4 in Microsoft Windows (XP, Vista or 7).
  + Used the Downloads feature of the Epi Info web page at [www.cdc.gov/epiinfo/](http://www.cdc.gov/epiinfo/) and
  + Chosen version 3.5.4 (NOT version 7 of Epi Info). Epi Info has not been tested under Windows 8, but it would probably work.
  + OpenEpi can be used directly from the internet, or from the files supplied for this course, and can also be downloaded from the DOWNLOAD feature on its menu.

TODAY, for your convenience, all the files are in the DVD or USB flash memory stick and we’re going to do all that installation stuff… right… **NOW!** Once that’s done, it doesn’t have to be done again.

Now, we’ll do some brief introductions of your partners in this course (Andy Dean and Consuelo Beck-Sague) and then…We’ll SHOW you how you can use Epi Info ANALYSIS to analyze an Excel file… That’s how we’ll introduce you to each other… at least, the 20+ who submitted the pre-course survey. We’ll also show you how you can use a “comma-delimited” file (which we made out of an Excel file) to analyze data originally from an Excel file.

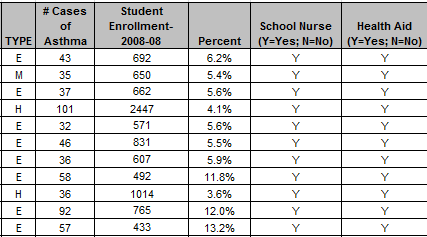
**“See one, do one, teach one”: See one…**

* Watch this—we’ll use the Excel file with the information of those participants that filled in the pre-evaluation to get to know each other.
* We’ll go to ANALYSIS
* We’ll use the command READ—it automatically defaults to Epi Info file, but we want it to read the Excel file that has your information… so we will arrow down to Excel 8…
* There… it found all the Excel files that we have; we’ll pick the Excel file called EpiInfoPre-course1
* We will use LIST to get a line listing of all the variable names and all of your answers…
* Look, using FREQUENCIES we can see what kind of computers you use.
* We had to go into Excel to change these variable names into shorter variable names that are easy to analyze and, after READing the file in Epi Info, we used the WRITE command to make this into an Epi Info file…
* Now, we can use MEANS to see the average and median number of years that you have worked in this field, and even compare the average number of years in the field by type of computer you use.
* In ANALYSIS, we made a new variable that collapses the 5 categories into beginner Epi Info user Y/N, and beginner Excel user Y/N, and we used TABLES to see if there’s any association.

**Asthma in Schoolchildren: Do one…**

* The data files for this exercise are supplied in a folder called “Epi Info Course FIU” which can be copied to your desktop or other location, or used from a USB memory stick. It can be copied from a CD or DVD, but the final location where it ends up must be “writable”.
* In 2010, a community-based organization collaborated with a school nurse organization and county schools, asking school nurses to provide the **number** of students in their schools whose medical forms (filled out by the students’ doctors) said that they had asthma. The nurses filled in the number of students who were enrolled in the school and the type of school (elementary, high school, etc.) and the zip code of schools.
* The results were entered into an Excel spreadsheet and edited to provide missing school addresses and zip codes…
* The Excel spreadsheet was used to export a text file with individual schools, each as a line of data (just like in the other spreadsheet, each of you was a line of data). The lines of data were separated by a carriage return (Enter) character and data items on a line separated by commas.

Excel Spreadsheet looked kind of like this: The comma separated value (CSV) file looks like this:



ZIP,TYPE,CASES, ENROLLMENT, SCHOOLNURSE,HEALTHAIDE

33010,E,34,702,Y,Y

33012,M,35,650,Y,Y

33012,E,49,742,Y,Y

33012,K8,6,1138,Y,Y

33012,E,65,740,Y,Y

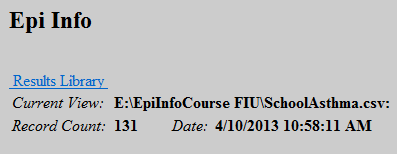
33012,M,49,1190,Y,Y

33013,E,55,827,Y,Y….. etc

Etc...etc… goes on for 130 schools

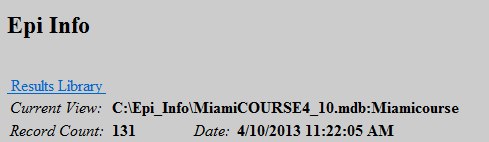
You saw before how Epi Info ANALYSIS was used to analyze an Excel file. Now, your mission (if you choose to accept it) is to summarize and analyze the data on asthma in schoolchildren by type of school, zip code, and other variables, and to make some graphic demos and generate hypotheses...

* Run Epi Info from its icon
* Click the button that says Analyze Data
* Choose READ(Import) from the list of commands on the left
* To READ(Import) our CSV file, choose in “Data Formats” Text (Delimited) from the menu
* Then for “Data Source”, got to “Removable Disk” or “Removable drive”…
* Then go to the bottom right, where it says “Text files” .txt—that’s a menu. You want to pick “Other Files (\*.csv, \*.tab, \*asc)”
* You’ll see a file “school asthma”
* Left click on open and it will bring it up--
* It will have the file there and the option to say “OK”—click on OK
* Say OK when it asks you stuff about the “file spec”, such as the first row containing field names.
* And you’re IN!! It should look like this:



That means you have 131 schools, and today’s date and time. Right above is the name and type of file that you are analyzing. You are ready to start asking questions… “the truth is out there”

* Find the “List” command on the left to get the lay of the land---use gridtable if you want to get all of them on one line, and html if you want to copy and paste onto a Word document.
* You don’t HAVE to make this into an Epi Info file—you could analyze it as a csv file… but let’s just do it… it’s easy! Go to the left column menu of commands again and pick “Write (Export)”
* On “output mode”, pick “replace”, in case you or someone else has already made the table that you intend.
* You have a choice of programs to which to export this… access… whatever… you want to pick “Epi2000”, which is the default. You have to look around for the others. Pick Epi2000.
* Where it says “File Name” put WHERE you want it to go (I want it to go to my Epi Info file in my C drive) and a “Data Table” (I called mine Miamicourse).
* The command “Write[export]” defaults to “all variables”; they’ll be clicked. But just make sure the “All(\*)” has a check mark by it.
* OK! You just created an Epi Info file in Microsoft Access Format, with the data!
* Now READ(Import). Click on “Change Project” if there’s one already picked out. Then it will give you all the Epi Info files you can get at. Pick yours.
* When it comes up, I click on Miamicourse
* Then I click OK. It should look like this—you have an Epi Info file.mdb, with Miamicourse



* OK, now you’ve got some explaining to do!
* Use the **FREQuencies** command in the left hand column in statistics to show you how many schools you have, by type of school.
* Use FREQuencies to show the number of students with asthma
* Use the Means command to show you the total number of students with asthma (“Cases”), the average and median, the range of number of students (“enrollment) per school.

## Discussion

What have we done so far, what is our overall analytic objective, and what would you like to do next? OK, so we imported and cleaned up the data and are moving now to analyzing and making sense of it.

**Defining and Assigning Values**

* To get at PREVALENCE of asthma per school, the proportion of students who meet the case definition of asthma, DEFINE a new variable. Pick “DEFINE” from the menu, and give the new variable the name AsthmaPct. Don’t bother to specify its type other than standard. You’re going to have a new empty column. You can check that it’s totally empty by doing List AsthmaPct.
* Now, you are going to assign a VALUE to that variable. Picking ASSIGN will bring up a little dialogue box. In the blank space where it says Assign Variable, pick out from “Available Variables”, a menu in the dialogue box “AsthmaPct”. Then below the words =Expression, you will:
  + put an open parenthesis, like this🡪 (
  + pick out from Available Variables Cases
  + pick out or just type / (for “divide by”)
  + pick out from Available Variables Enrollment
  + Close parentheses 🡪 ) then put “\*100” so it will express it as a percent.
* OK, now do a MEANS of your Asthmapct variable. In the summary below, what’s the mean asthma prevalence? The median? The minimum? The maximum?
* Now, this time, let’s compare types asthma prevalence by type of school. Pick MEANS from the left side menu, and pick from the variables Asthmapct for the “MEANS for” blank, and in “Cross-Tabulate by value of” put“Type”, that’s the type of school.
* Click OK

**The p-value at the very bottom of the output tells us whether or not there are differences among the columns. If it is less than 0.05, there are indeed “significant” differences. If class time permits ask the instructor how to combine the columns and come up with means for the younger and older students.**

* What type of schools has the highest median asthma prevalence? What type has the highest maximum asthma prevalence? Which type have the lowest?
* To collapse all school types into the school type with the highest prevalence and all the others, you can define a new variable using DEFINE.
* The, instead of using “ASSUGN” pick IF from the left column. Under IF put Type=”E”, now go to “Then” which will bring up the menu on the left—NOW pick “ASSUGN Variable” and type in “Elementary”. Go below to =Expression and just pick out from the expressions “Yes” and left click on the button “Add”.
* Click on “Else” (that’s for all other types of schools) pick “ASSUGN Variable” and get Elementary and pick out the expression “No” and left click on the button “Add”. Now, click on OK—
* You can now do MEANS Asthmapct and cross-tabulate by value of “Elementary” the new Yes/No variable.
* Look at the comparison now. Compare the medians of Elementary vs other schools. Look at the very lowest p-value that’s like a rank sum test. Is the difference in asthma prevalence in these Elementary schools significant?

## Say we now want to format numbers and convert Numbers to text and back again

1. Let’s fix the long decimals. DEFINE another variable called RoundPct. You can use the ASSIGN command again and ASSIGN RoundPct variable to one decimal place. To find out how, click on HELP in the ASSIGN dialog, and then find the documentation for the FORMAT command by clicking on the Text function item.. Hint: the format “##.#” will probably work, since our percentages are never as large as 100. See if RoundPct is simpler by doing a LIST. You can LIST one or several variables by selecting them in the LIST dialog.
2. The problem with FORMAT is that it produces a text result—good for display, but not for calculations. We can convert the text items back to numbers with a function called TXTTONUM. (Note the missing E and the two T’s). DEFINE a variable called RoundNum and ASSIGN it the value of TXTTONUM(RoundPct). Check the results with LIST.

## Saving Commands in a Program or PGM

1. Things are getting more complicated, and this would be a good time to notice that all the commands you have given are recorded in the program (pgm) that is developing in the lower right panel--the program editor. Click the RUN button in the program editor. This will run the entire program and repeat all the steps you have performed so far, including the importation of the csv file, the frequencies and LISTs. If there are mistakes or unwanted commands like extra LISTs, you can correct the program in the editor.
2. To save the program for later use, use the SAVE button in the Program Editor, and specify a name for the program as “AsthmaByZip”. This will save the program in the database, from which it can be loaded and run with the OPEN button. You can also save the program with the extension PGM outside the database by using the SAVE AS A TEXT FILE option in the SAVE dialog.
3. To demonstrate the value of saving the program, erase the current program from the Program Editor by clicking the NEW button. Now use the OPEN button to find your saved program and load it into the Program Editor. Use the RUN button to run it again, as before. Note that you can also load and run the program from the RUN SAVED PROGRAM command in the left panel.

## Frequencies and MEANS

In the FREQuency Command, choose the AsthmaPct variable and click OK. The result tells us how many times each value of AsthmaPct is in the database, but most of the counts are 1, since we have not grouped the percentage data, and it is expressed to the nearest 0.1%. We will fix this later, but, for now, let’s try MEANS with AsthmaPct. This is a bit more informative, since MEANS works well with “Continuous” data (quantitative measurements like height, weight, or percents), while FREQuencies are intended for “Categorical” data like city names, or descriptive categories like “Red”, “Green”, and “Blue”. What variable do we have in our database that will best be described by a FREQuency? (Hint: How many high schools do we have?)

Run the MEANS command again—this time by placing the mouse cursor on the MEANS AsthmaPct that has already appeared in the Program Editor and clicking on RUN THIS COMMAND. This is a convenient way of running a single line of code, perhaps after making changes or cleaning the data. Starting with examination of the MEANS result at the bottom of the output page, and then using other commands as needed, answer the following questions:

1. What is the Mean (average) frequency of Asthma registrations in the schools under study? The Median (50th percentile)? If they are different, which is the most useful?
2. What are the minimum and maximum Percents?

## Missing Values

What is happening to the Missing values for AsthmaPct, and how many are there? Click on the SET command in the left panel. Review the possible settings and then click on INCLUDE MISSING before repeating FREQ AsthmaPct with RUN THIS COMMAND, as above.

## Managing Output

The upper panel on the right is really a browser. The output of Analysis is constructed dynamically in web page (HTML) format. The name of the HTM file is at the top of the screen (probably “OUTxxx.HTM”). You can use the ROUTEOUT command to use a different name, like AsthmaResults.HTM, for subsequent output. Use ROUTEOUT to specify a name. An easy way to repeat previous commands is to place the cursor on a command and then click the RUN THIS COMMAND button. The output that is produced will now go to AsthmaResults.HTM, from which it can be displayed in a browser or edited in Microsoft Word or OpenOffice.

## Graphing Data

Use the GRAPH command to make a Histogram of AsthmaPct to see if the values are in the shape of a “normal” curve (are normally distributed). Normal distributions have a big hump in the middle, tailing off toward both ends, and are more or less symmetrical.

Since we have four different types of schools, GRAPH can show the Average of AsthmaPct as a Bar graph with a bar for each TYPE of school.

Now that you know how to make a Histogram of AsthmaPct, try instructing GRAPH to make a separate graph for each value of school TYPE. Since we know there will be 4 TYPES, set the page display for 2 horizontal and 2 vertical graphs.

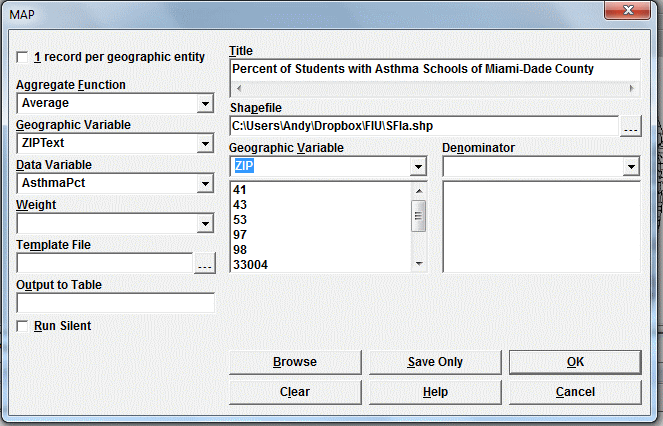
## Mapping the Data

Mapping requires a suitable “shapefile” of the area. These files can be downloaded from many websites, and those for states and counties as well as other countries are available from the CDC EpiInfo website ([www.cdc.gov/epiinfo/](http://www.cdc.gov/epiinfo/)). We will be using a shapefile of South Florida that was prepared from a larger map of Florida ZIP codes.

To display the shapefile go to the main Epi Info menu and click the button on the right that says CREATE MAPS. This runs the EpiMap program. In the FILE menu, click MAP MANAGER. In the manager, choose ADD LAYER, and then navigate to SFla.SHP. You should see a detailed map of South Florida ZIP codes. Close EpiMap and open the Analysis program from the main menu, since we will be sending our data to the map through Analysis.

In order to have the shapefile display the values of AsthmaPct that we have in our dataset, the dataset must have a variable that exactly matches a variable in the shapefile. . The Epi Map program will match the ZIP values and display AsthmaPct in the corresponding polygon on the map as a color or shade density.

In this case there is a ZIP variable in the shapefile and a ZIP variable in our dataset, but it turns out that the dataset ZIP is numeric, and the names of polygons in a shapefile are always text variables. Use the DISPLAY command to discover ZIP’s type in our dataset, and then DEFINE a new variable called ZIPText. ASSIGN ZIPText the value of FORMAT(ZIP) to convert it to text. Now you can click on the MAP command and fill in the dialog as follows:



Why did we choose the AVERAGE value of AsthmaPct to display rather than “One value per geographic entity”?

After you have made the choices and told MAP to match the ZIPText values to the ZIP values in the shapefile, click on OK. After a pause, you should see a scary message about the number of ZIP codes for which there is no data. Just click CONTINUE, and you should see the map with shades of blue representing the values of AsthmaPct. A legend appears to explain the different values. With the mouse, move the legend until it is close to Miami-Dade. Since we want to display only the Miami-Dade area, click the button at the top of the screen that looks like a magnifying glass. Use the mouse to click and drag a box around the desired map area until it fills the screen.

If you prefer a different color or pattern of shading, experiment with the MAP TYPE | Choropleth entry at the top of the screen. This offers many opportunities for customization.

Exit from the Choropleth dialog and choose DOT MAP from the MAP TYPE menu entry. This will make a dot for each percentage point represented, an entirely inappropriate way to display percentages, but a good way to show individual cases in a surveillance system, for example. The Choropleth map is definitely the right choice for our AsthmaPct data.

This completes the Analysis exercise…….

What have we done, and how might it be applied to other health-related challenges?