Analysis

Introduction

In Epi Info™ 7, Classic Analysis acts as a statistical toolbox, providing many ways to transform data and perform statistical evaluation. Data can be selected, sorted, listed, or manipulated with a series of commands, functions, and operators. Available statistics include frequencies, means, and more advanced processes (i.e., Kaplan-Meier survival analysis and Logistic Regression).

It reads data files created in Form Designer and other types of databases (e.g., MS Access, MS Excel, SQL Server, and ASCII). Classic Analysis can also produce graphs to present graphic representations of data.

Click **Classic** from the Epi Info™ main window or select **Tools > Analyze Data>Classic** to analyze data. Note that the Classic Analysis module contains four areas: the Command Tree, Program Editor, Output window, and the Message Area.

1. The Command Tree contains a list of available commands separated into folders by command type on the left of the window. Selecting a command opens the corresponding dialog box for that command, function, or statistic to run.
2. The Program Editor displays the commands and code created using the Classic Analysis Command Tree. Commands can also be typed directly into the Program Editor. Programs or .PGM files written in Classic Analysis can be stored in the current .PRJ or as text files. Saved programs can be run against new data, or shared with others.
3. The Output window above the Program Editor acts as a browser and displays information generated from commands run in Program Editor. The buttons allow you to navigate through program scripts that run and display in the output screen.
4. The Message window alerts you if any problems occur with any executed commands.

Read an Epi Info 7 Project

In order to analyze data, you must read or import it into the Classic Analysis module. The **Read** command allows you to select a project and/or data table to run statistics.

1. From the Classic Analysis Command Tree Data folder, click Read. The READ dialog box opens.
2. The **Database Type** field indicates the database file to be loaded Epi Info 7 Project, Microsoft Access, Microsoft Excel, Microsoft SQL Server Database, MySQL, Flat ASCII file). Select Epi Info 7 Project.
3. The **Data Source** field indicates the file location/path. Click on the **ellipsis (…)** button on the right of the Data Source field to open the **Select a data source** dialog. Locate the *RMS Course* folder, and open the *AsthmaSurvey2005.prj* file.
4. The Data source explorer shows the available Forms and Tables in the project. Select the *PreInterventionSurvey* form. Click **OK.**
5. The READ command is saved in the Program Editor and run simultaneously. The current form file location, Record Count, and Date appear in the Classic Analysis Output window; the READ command appears in the Program Editor.
    - *How many records are in the AsthmaSurvey2005 dataset?*

The List Command

The List command creates a line listing of the current dataset. List can be customized to list all, exclude or show specific records. There are 2 modes for displaying data listed under the Display Mode section of the dialog box:

1. Printable/Exportable
2. Grid

The Printable/Exportable mode differs from the Grid in that rather than being spreadsheet-like in appearance, the data are written to the Output screen in a web table format, which is useful for printing the data. Web display mode creates an embedded output file.

- From the Command Tree, click Statistics > List. The LIST dialog box opens. Click **OK** to accept the default settings.
- The variables in the dataset appear in a grid table inside the Output Window. To navigate LIST * GRIDTABLE results, use the Tab key to move forward one cell at a time and the Shift+Tab keys to move back one cell at a time. Navigate through the records and cells using the left, right, up, and down arrow keys.
- Again click **List**>select **Printable/Exportable>OK**

The Sort Command

The SORT command specifies the sequence in which records will appear when using the List, Graph or Write commands. Sort organizes the listed data in an ascending or descending order, based on selected variables.

For example, you can sort by last name or age.

- Click **Sort**> double click **DOB**>**OK**.
- Use the **List** command to view all the variables>Cancel Sort>OK

You can also create a list of sorted variables.

- Click Sort>double click **ChildLast**>**OK**
- Click **List**>**Printable/Exportable**. From the Variable drop-down, select **ChildFirst**, **ChildLast, Asthmatic Conditions**>**OK**
- Cancel Sort>**OK**.

The SORT command specifies the sequence in which records will appear when using the List, Graph or Write commands. Sort organizes the listed data in an ascending or descending order, based on selected variables

The Select Command

Sometimes you may want to select a subset of records for analysis. For example, you might want to perform some analyses only for males, or those <9 years of age. This can be done using the Select command. Clicking of the Select command in the Analysis command window will lead you into a Select command dialog box.

- Click **Select**. From Available drop-down, select Gender. Click the = operator. **Click "**, and type **Male**. Click **"** to close the expression. Click **OK**. The word Male must be enclosed in

quotes because it is a text field.

- The program Editor contains the Select command information.
- List the selected records. (**List > Printable/Exportable > OK**)
- How many male students are aged less than 9 years?
- Always cancel Select and Sort at the end.

Display

This command displays table, form, and database information. With it, you can see all the variables in the dataset, including names, field types, and format information. It is important to use display prior to merging or creating statistics to ensure that field types and variables names have been coded as needed.

From the command tree window click **Variables** from the Display folder. Click **OK**. The Variables in the *AsthmaSurvey2005* project appear in the Output window.

Set

This command provides various options that affect the performance and output of data in Classic Analysis. These settings are utilized whenever the Classic Analysis program is used.

From the command tree Options folder, click Set. *Set the following options:*

1. Yes as Yes
2. No as No
3. Missing as Missing
4. Select Show Complete Prompt. Others are grayed out in this version of Epi Info 7. They will be available in future versions.
5. Leave Include Missing as not selected. We don't want missing values to distort the analysis.
6. Process Undeleted Records only (Normal)
7. Click **OK**

The Merge Command

This command can be used to join information from a Microsoft Excel spreadsheet to an Epi Info 7 project file. In this exercise, you will use the Merge command to join the 12 records you just created when entering data with a larger data table gathered using a Microsoft Excel spreadsheet. A merge requires a Build Key that represents a matching variable inside both sets of data. You will use the *Build Key* StudentID.

1. If you haven't already done so, Read the *AsthmaSurvey2005* project. You should have 12 records.
2. Click Merge. From the Data Formats drop-down, select **Microsoft Excel 97-2003 Workbook (.xls)**
3. Click the Data Source Browse button. The Open Existing File dialog opens. Browse to the

AsthmaSurvey2005 subfolder in the *RMS Course* Folder on your C drive. Select the file called **AsthmaTable.xls** > **Open**

4. Click the Worksheet called **SurveyData$**. The SurveyData$ is the named worksheet in Excel file. Select both Update and Append. Click **Build Key**.
5. The Build Key Dialog box opens. From the list of **Current Data Variables**, select **StudentID**. From the **Related Tables Variables** list, select **StudentID.** Click the **Add Relation** button. You will notice that **StudentID = StudentID** appears in the Key Component section.
6. Click **OK**. Then Click **OK**. The merge process should take a while (*now would be a good time to take a 2-3 minutes break*). The status bar will show a message - *merge in-progress: xx of 800 records processed…*
7. The Output window will give you a summary of the number of records updated (*How many?*) and the number inserted (*How many?*)
8. List the variables in Grid. Notice the first 12 records contain information from your survey. The StudentIDs are the same, so the records were updated to reflect any changes. Use the Close X to exit the grid listing and return to the Output window. Notice the Program Editor shows the commands used to create the Merge and the List.

Frequencies

The frequencies command produces a frequency table that shows how many records gave a value for each variable, the percentage of the total and a cumulative percentage.

1. Click **Frequencies** under the Statistics folder of the Analysis Command Tree.
2. From the **Frequency of** drop-down, select **SchoolNum**. Click **OK**.
3. Frequency can be stratified or grouped by a selected variable.

- For example, to determine which school had the highest frequency of students with bronchitis,
- Click Frequencies. From the Frequency of drop-down, select **Bronchitis**.
- From the **Stratify By** drop-down, select **SchoolNum** >**OK**

You can use the SELECT command to specify a subset of students and then create a frequency for only those records. For instance, you may be interested in the number of students who have Reactive Airway Disease and if they are currently taking medication for the condition.

1. Use the select command to view those who answered Yes to having Reactive Airway Disease. (Hint: RAD=(Yes)
2. Create a Frequency of students with RAD who have been prescribed medication (Hint: Frequency of Medication)
3. Cancel Select to return to the original 800 records
4. *What other way can you use to find the answer?*

MEANS

The Means command is used to compute descriptive statistics for a continuous numeric variable. When used with a cross-tabulation variable, it also computes statistics showing the likelihood that the means of the groups are equal. The mean of a YES/NO variable is the proportion of respondents answering yes.

Means can be created for continuous or categorical values. Continuous variables fall along a continuum with one variable selected, for example, age. Categorical variables are groupings or categories with two variables selected, for example, gender. If one variable is selected, the program produces a table of descriptive statistics.

If two variables are selected, the first is a numeric variable containing data to be analyzed and the second is a variable that indicates how the group will be distinguished. The output of this format is a table of descriptive statistics of the numeric variable for each value of the group variable. Means produces the following statistical tests:

Parametric test	***Non-parametric test***
Student T-test	Mann-Whitney/Wilcoxon Two-Sample test
ANOVA for 2 or more samples	Kruskal-Wallis one-way ANOVA for 2 or more samples
	Bartlett's test for inequality of Population variance

1. To determine the mean no of missed days due to Asthma, click **Means**. From the **Means of** drop-down, select the variable **MissDays.** Click **OK**.
   - *What is the mean no of days missed due to Asthma?*
2. Means Command can also be used with a cross-tabulated variable. For instance, to determine which school had the highest average of missed days, *Click* Means > *Means of*...**MissDays** > *Cross tabulate by Value of...* **SchoolNum** > **OK**.
3. You can also create a Stratified Means. For instance, to determine if the school with the highest average of missed days for males is the same as for the females, Click Means > *Means of*...**MissDays** > *Cross-tabulate by Value of*...**SchoolNum** > *Stratify by*...**Gender** > **OK**.

You can use Select with the Means Command. For example, you may want to determine if the average number of missed days is higher among those who answered Yes to asthma than the overall average. To do so, Calculate the Mean number of days missed due to asthmatic conditions. Then, click select. From the Available Variable drop-down, select Asthma. Use the Function and Operators buttons to create the code Asthma=(yes) in the Select Criteria field. >OK. Create a Means of Missdays. Compare the two results. Click Cancel Select.

Date calculation

List the Age variable in a grid. What do you notice?

Click Assign. From the Assign Variables drop-down, select Age. In the =Expression field type your code. `Age=YEARS(DOB, 25/10/2005)`. Click OK. Now list the Age variable in a grid. Notice what has happened. Calculate the Means of Age.

Tables

Tables are used to create cross-tabulations of categorical variables. Tables can help to determine whether a risk factor is linked to an outcome. For 2x2 tables, the command produces odds ratios and risk ratios. For tables other than 2x2, Chi-Squared statistics are computed. You

may want to answer the question: *Does one school have a higher number of students with asthma*?

1. Click **Tables**. From the Exposure Variable drop-down select SchoolNum. From the Outcome Variable drop-down, select Asthma > OK.
2. *How many students had both Asthma and Wheezing*?
3. Click Tables. Select the Exposure Variable of Asthma. Select the Outcome Variable of Wheezing. OK

Saving a Program

Click **File > Save Pgm**. A Save Program dialog box opens. In the program Field, type Analysis Lesson. In the Author field, type your initials. In the comments, type *Research Methods Course 2014* > OK.

To test the Saved Program, open Analysis. From Program editor, click File > Open Pgm. The Open Program dialog box opens. Select the *AsthmaSurvey2005.prj* file in the Project File field. From the Program drop down, select Analysis Lesson. Click OK. Your saved program appears in the program editor. Select Edit > Program beginning. You can scroll to any previous command created, Highlight the command and click Run Commands. The results of running the program will appear in the Output window.

Defining New Variables and Recoding Data

A common task when analyzing data is to create or DEFINE new variable. These new variables may be a categorization of existing variable e.g. age to age group or a calculated field e.g. BMI from Weight and Height. Recode Command is used to place existing values in a new order.

1. Click Define. The define dialog box opens. In the Variable Name field, type *MissDays2*. >OK
2. Click Recode. The RECODE window opens.
3. Use the **From** drop-down box to select *MissDays*. You are taking the values from the variable *MissDays* and recoding them to appear in the variable *MissDays2*. We want to categorize the days missed due to asthma.
4. From the **To** drop-down box, select *MissDays2*.
5. In the **Value** (blank=other) field type “0” and press Tab.
6. In the **To Value** (if any) field type “0” and press Tab.
7. In the **Recoded Value** field type “No Days Missed” and press Tab. A new blank entry line appears.
8. Complete the code as follows.
   - Value “1” To Value “5” = “1 - 5”
   - Value “6” To Value “10” = “6 - 10”
   - Value “11” To Value “15” = “11 - 15”
9. Click OK.
10. Frequencies > *Frequency of*...**MissDays2** > OK

IF statement

IF/THEN has been used to create a skip pattern in Check Codes. IF/THEN statement can also be used in analysis to do most of what Recode can do. For instance, if you regard any student who missed less than 12 school days in the pre-intervention survey as having good attendance, you can use the If/Then command to identify the students with good attendance. Read the pre-intervention survey.

1. Define a new variable **GoodAttendance** > OK.
2. Click If. Use the dropdown arrow to select MissDays, Click the < button. Type 12.
3. Click on the Code Snippet button in the Then field. Select **Variables > Assign**.
4. The Assign Dialog box appears. From the Assign Variable drop-down, select *GoodAttendance*. Place the cursor in the =Expression field. Click Yes. Click OK.
5. Click on the Code Snippet button in the **Else** field. Select **Variables > Assign**. From the Assign Variable drop-down, select *GoodAttendance*. Place the cursor in the =Expression field. Click **No**. Click OK.
6. The code appears in the program editor. Click OK.
7. Now Click **Frequencies** > **Frequency of**...**GoodAttendance**.
8. You can Select GoodAttendance = Yes. You can analyze the selected group as you wish.

Use Related Data

To use RELATE, the READ command must open at least one form/table. The form/table to be linked requires a key field that relates records in the two forms/tables and can be used to join them together. The keys in the main and related tables or forms do not require the same name. If the table was created in Classic Analysis and the data entry completed using Enter, Classic Analysis can establish a relationship automatically using the Global Record ID and Foreign Key variables created by Epi Info 7. If the relationship was created in another program that uses different keys, the key variables in both files must be identified. Relationships are represented by double colons (::). Classic Analysis can establish relationships using multiple keys.

After issuing the RELATE command, the variables in the related table may be used as if they were part of the main table. If variable names are duplicated in the related tables; variable names will be suffixed with a sequence number. Frequencies, cross-tabulations, and other operations involving data in the main and related tables can be performed. The WRITE command can create a new table containing both sets of data. More than one table can be related to the main table by using a series of RELATE commands.

Relate keys can contain mathematical or string concatenation expressions. Epi Info 7 functions can be used to determine relationships. It may be easier, however, to write out a new file in which the complex key is included as a single variable. Use this single variable as a key.

Example

The **Sample.prj** file contains two forms: **Surveillance** and **RHepatitis** which are related. The variable GLOBAL RECORD ID is the internal Epi Info 7 identification key located in more than one form in the project.

1. Read in the Sample.PRJ project from the *C:\RMS Course\Projects* folder. Open Surveillance.

2. Click RELATE. The RELATE dialog box opens.
3. Select **RHepatitis**. The Build Key button activates.
4. Click Build Key. The RELATE-BUILD KEY dialog box opens.
5. From the Current Data Variables list, select **GlobalRecordId**.
6. From the Related Table Variables drop-down list, select **FKEY**.
7. Click the **Add Relation** button. The **Key Component** field populates.
8. Click OK. The RELATE dialog box opens and the Key field populates with the join information: GlobalRecordId :: [FKEY]. Click **OK**. *How many records?*
9. List

Output Files

Each time you read a file into Analysis and run a command, you create an Output file. The Output appears in the Output window and in the Output file inside the default Epi Info 7 Output directory. When an output file is created, the name and location of the file appears in the Output window title bar as a reference. Outputs are added to the same file until the data source is changed.

From the Output window button bar, click Open. Select one of the OUTPUT files. Click Open. The file opens in your Output window.