Epi Info 2000 Basics

Epi Info 2000 (EI2K) is an epidemiologic data management and analysis program written and supported by the Centers for Disease Control and Prevention. The previous version of Epi Info (version 6) ran on DOS computers and enjoyed wide distribution (over 100,000 users world-wide; almost universal use in Health Departments). The new version is a Windows program written in Microsoft VisualBasic, and is still feeling its way.

Data Entry and Documentation

Epi Info 2000 uses two key programs for data entry. These are:

- MakeView.exe
- Enter.exe.

Briefly, constructing a data based is a two step process. First, you use the *MakeView* program to define your variables and construct your data entry screen(s). Then, you use the *Enter* program to enter your data. A good way to get acquainted with these processes is to take the *Brief Tour* that starts on page 29 of the *Epi Info 2000* Manual.

The issue of accurate data entry is of course crucial. Unfortunately, the current version of Epi Info does *not* support double entry and validation procedures. It does, however, provide for range checks and automatic labeling.

Once data have been entered, it is good form to document your data in the form of a *codebook*. Codebooks contain information to help the user decipher the contents and structure of the data file. In general, the codebook should include information about the file itself (e.g., filename, number of records, file location, storage media, file type, dates of creation and modification) and information about the variables in the file (e.g., variable names, questions from the survey that elicited the response, units of measure, ranges, coding schemes, descriptive statistics for variables).

Data Backup and Security

To paraphrase a well-known computing saying, “There are two kinds of computer users. Those that have lost a major chunk of data, and those who are going to lose a major chunk of data.” Since disasters such as fire, theft, and earthquakes do happen, it is best to backup all elements of a project (e.g., data files, code books, software, software settings, computer programs, word processing documents, etc.) and keep backed-up media off-site at a separate location. Acceptable backup media include floppy disks, Zip disks, tape, CD-ROMs, and outside networks.

Backup procedures should be thoroughly tested to ensure that archived files are uncorrupted and can be easily restored. Procedures should be written up so that so that personnel unfamiliar with backup and restore methods could follow the procedure if necessary. The entire process should be kept as simple as possible.

Epidemiologists and other health researchers need to be aware of the ethics of working with the private nature of research files. This is especially important when data contain personal identifiers and confidential medical information. It is each researcher's duty to make him or herself aware of local, national, and international laws governing confidentiality and follow ethical guidelines governing use of their data. Researchers must make it their responsibility to protect people's privacy.

Many confidentiality concerns can be allayed by using anonymous data files (i.e., data on individuals but without personal identifiers). On the other hand, it is not always clear when data are anonymous. For example, when working with a small population or with a rare event, it may be possible to identify individuals even if their identity is not explicit. In such instances, it is not clear how far the epidemiologist's responsibility extends in protecting identities.

Introduction to the Analysis Program

Epi Info's ANALYSIS program manages, prints, and performs statistical analyses on EpiInfo data (MDB) files. Instruction on the use of Analysis appear on 31 - 45 of the manual. Briefly, ANALYSIS program has three windows. These are (a) a command choice window (on the left), (b) an output window on the top (which is a simple html a browser), and (c) a command window, on the bottom of the screen.

Commands are selected by clicking on the commands in the command window on the left. This brings up dialogue a box that help construct text commands. The text command appear directly on the bottom screen.

To start a analysis session, you must first READ the data set into the session. Make certain you click the Change Project button before selecting a new file, or the current data file could be inadvertently linked to the current file. Then, LIST the data to make certain the data has been read in. There are three LIST options: listing the data to HTML output, listing to a table that doesn't allow modification, and listing to a table that does allow modification to the data. I suggest using the middle option in most instances.

Next, at your option, use the ROUTEOUT command (under the OUTPUT folder in the left-hand window) to direct future output to a file with a given name. Use logical names to direct your output. For example, when working data called hdur.MDB, you might ROUTEOUT to hdur-output. The output will be a typical .html file (with the .HTM extension).

Now you are ready to run some statistics. Before proceeding with more complex analysis, it is usually wise to get descriptive statistics for each variable. Use the MEANS command to get descriptive statistics for continuous (quantitative) variables, and use the FREQ command to get descriptive statistics for categorical (qualitative) variables.

Exercises

(1) HDUR: *Hospital Duration Stays* (Townsend et al., 1979; Rosner, 1990, p. 36) Data represent a sample from a hospital discharge study. Data represent:

- DURation of hospitalization (days)
- AGE (in years)
- SEX (M/F)
- Body TEMPerature (degrees Fahrenheit)
- White blood cell count (x 100/dl)
- In-hospital antibiotic use (AB
- Whether a blood CULTure was taken (1 = yes, 2 = no)
- Admitting SERVice (1 = medical, 2 = surgical).

DUR	AGE	SEX	TEMP	WBC	AB	CULT	SERV
5	30	F	99.0	8	N	2	1
10	73	F	98.0	5	N	1	1
6	40	F	99.0	12	N	2	2
11	47	F	98.2	4	N	2	2
5	25	F	98.5	11	N	2	2
14	82	M	96.8	6	Y	2	2
30	60	M	99.5	8	Y	1	1
11	56	F	98.6	7	N	2	1
17	43	F	98.0	7	N	2	1
3	50	M	98.0	12	N	1	2
9	59	F	97.6	7	N	1	1
3	4	M	97.8	3	N	2	2
8	22	F	99.5	11	Y	2	2
8	33	F	98.4	14	Y	1	2
5	20	F	98.4	11	N	1	2
5	32	M	99.0	9	N	2	2
7	36	M	99.2	6	Y	2	2
4	69	M	98.0	6	N	2	2
3	47	M	97.0	5	Y	2	1
7	22	M	98.2	6	N	2	2
9	11	M	98.2	10	N	2	2
11	19	M	98.6	14	Y	2	2
11	67	F	97.6	4	N	2	1
9	43	F	98.6	5	N	2	2
4	41	F	98.0	5	N	2	1

(A) Create an EI2K data file with these data. Name the data base Hdur.MDB and name the data entry screen (view) Hdur.

(B) Compute frequencies for the categorical variables in the data set and compute means, standard deviations, minimum, and maximums for each continuous variables.

(C) Document the data set in the form of a DD file.

(2) WCGS: *Western Collaborative Group Study* (Selvin, 1991, p. 41). Data from a study on type A behavior and cholesterol are:

ID	CHOL	BEHAV
1	233	A
2	291	A
3	312	A
4	250	A
5	246	A
6	197	A
7	268	A
8	224	A
9	239	A
10	239	A
11	254	A
12	276	A
13	234	A
14	181	A
15	248	A
16	252	A
17	202	A
18	218	A
19	212	A
20	325	A
21	344	B
22	185	B
23	263	B
24	246	B
25	224	B
26	212	B
27	188	B
28	250	B
29	148	B
30	169	B
31	226	B
32	175	B
33	242	B
34	252	B
35	153	B
36	183	B
37	137	B
38	202	B
39	194	B
40	213	B

Create an EI2K file with these data, report routine summary statistics and frequencies, and document the data in the form of a code book.

(3) %IDEAL: *Diabetic Body Weight* (Pagano & Gauvreau, 1993). Data representing body weight expressed as a percentage of each subject's ideal body weight are: 107, 119, 99, 114, 120, 104, 88, 114, 124, 116, 101, 121, 152, 100, 125, 114, 95, 117. Store these data in an EI2K file.

(4) TOX-SAMP: A study of cerebellar toxicity associated with a chemotherapeutic agent was completed through joint collaboration between the US Food and Drug Administration and the University of Wisconsin teaching hospital. Some of the data from the project is listed in the table below. Variable include information on patient identification (ID), patient age (in years), sex of the patient (1 = male, 2 = female), manufacturer of the drug (S = Smith Co., J = Jones Co.), entering DIAGnosis (1 = leukemia, 2 = lymphoma), STAGE of disease (1 = relapse, 2 = remission), TOXICity (1 = yes, 2 = no), DOSE of drug (gms/M^2), serum creatinine level (SCR, mg/dl) and body WEIGHT (in kilograms). Data are:

ID	AGE	SEX	MANUF	DIAG	STAGE	TOX	DOSE	SCR	WEIGHT
1	50	1	J	1	1	1	36.0	0.8	66
2	21	1	J	1	2	2	29.0	1.1	68
3	35	1	J	2	2	2	16.2	0.7	97
4	49	2	S	1	1	2	29.0	0.8	83
5	38	1	J	2	2	1	16.2	1.4	97
6	42	1	S	2	2	2	18.0	1.0	82
7	17	1	J	1	2	2	17.4	1.0	64
8	20	1	S	2	2	2	17.4	1.0	73
9	49	2	J	1	1	2	37.2	0.7	103
10	41	2	J	1	2	2	18.6	0.9	58
11	20	1	S	2	2	2	18.0	1.1	113
12	55	1	S	1	1	2	36.0	0.8	87
13	44	2	J	1	1	1	22.4	1.2	59
14	23	1	S	2	2	2	39.6	0.8	83
15	64	2	S	1	1	2	30.0	0.9	69
16	65	1	S	1	1	1	23.2	1.7	106
17	23	2	S	1	2	2	16.8	0.9	66
18	44	1	S	1	2	2	17.4	1.0	84
19	29	2	S	2	1	2	18.0	0.7	56
20	32	1	S	1	2	2	18.0	1.0	84
21	18	1	S	2	2	2	17.4	0.9	70
22	22	1	S	1	1	1	26.1	1.7	69
23	43	2	J	2	2	2	18.0	0.8	63
24	39	2	S	1	2	2	18.0	0.9	55
25	38	2	J	1	1	1	16.0	1.0	112

Enter these data into an EI2K data file and save these data for future analysis. Include a code book on your disk (see p. 2 on how to set up a DD file).