EpiAnalysisTour - an introduction to **EpiData Analysis**

Http://www.epidata.dk

Version 25th. August 2005

EpiData Analysis

EpiData Analysis is used for:

- Basic descriptive Analysis of quantitive data
- Defining and modifications of data
- Editing / correcting data already entered
- Graphing Data
- Asserting that the data are consistent across variables
- Printing or listing data for documentation of error-checking and error-tracking
- Writing datafiles

EpiData Analysis works on Windows 95/98/NT/Professional/2000/XP and Machintosh with RealPc emulator. Linux based on WINE.

Suggested citation of EpiData Analysis program:

General reference for Analysis and EpiData DataEntry is:

> Lauritsen JM. (Editor) EpiData Data Entry, Data Management and basic Statistical Analysis System. Odense Denmark, EpiData Association, 2000-2005. (Available from Http://www.epidata.dk).

EpiData Analysis is based on a number of individual contributions:

> Mahmud S. Design and Implementation of core parsing principles and modules for EpiData Analysis. 2003 (source code).
>
> Lauritsen JM. Output layout definition in EpiData Analysis v1.2. EpiData Association, Odense Denmark, 2005. Http:\\www.epidata.dk\analysisinfo\html output definition.htm
>
> Development and Specification of SPC modules in collaboration with Gruk, Norway: V.Høgli (2004), B. Nyar (2004-)
>
> Testdesign and documentation: J. Hockin(2004-)

Contributors (period involved):

> JM.Lauritsen (2001-) Coordinator. Design, interface, implementation, documentation, testdesign, programming.
>
> S.Mahmud (2002-2003). Design, Programming and Implementation of core functionality.
>
> Programming and specification: M.Bruus(2004-). Programming: T.Christiansen(2004-)

Other Contributors

Specific testing: Several users (HB.Rieder, Neville Verlander, V.Høgli, B. Nyar, C Green, Jamie Hockin, Pedro Arias, Louk Meertens and others)

Statistical methods: S Kreiner (2004-)

Isolated parts of source code based on freeware and shareware components: See Http://www.epidata.dk/credit.htm

Suggested citation of EpiAnalysisTour introduction:

Lauritsen JM. EpiAnalysisTour - An introduction to analysis of quantitative data by use of EpiData Analysis. The EpiData Association, Odense Denmark, 2005.
Http://www.epidata.dk/downloads/epianalysistour.pdf (See Version above)

For further information and download of latest version: See http://www.epidata.dk

References for other parts of the whole EpiData system:(incomplete)

Lauritsen JM & Bruus M. EpiData (version 3). A comprehensive tool for validated entry and documentation of data. The EpiData Association, Odense Denmark, 2003-2005.

Bruus M. Implementation of chk and rec file structures in EpiData Analysis. 2005.

Bruus M. File Structure and Definition in EpiData. Http://www.epidata.dk.

EpiData and EpiData analysis are extensions based on the commands and standards defined in the Epi Info v6 software for DOS originally published as

Dean AG, Dean JA, Coulombier D, Brendel KA, Smith DC, Burton AH, Dicker RC, Sullivan K, Fagan RF, Arner TG, Epi Info, Version 6: A Word-Processing, Database, and Statistics Program for Public Health on IBM-compatible Microcomputers, Centers for Disease Control and Prevention, Atlanta, Georgia, U.S.A., 1995.

Parts of analysis has been based on rewritten source code extracts from the Epi6 source code into analysis (MLE and stratified analysis).

Funding Please acknowledge contributors presented at: Http://www.epidata.dk/funding.htm

Modfication of this document:

See general statement on www.EpiData.dk. Modified or translated versions must be released at no cost from a web page and a copy sent to info@epidata.dk. Frontpage cannot be changed except for addition of revisor or translator name and institution.

Disclaimer

We made every possible effort in producing a fail-safe program, but we cannot in any circumstance be held responsible for error, loss of data, work time or other losses incurred by or in relation to the program. Assurance of correct estimations and results lies with the end-user.

No pay for EpiData

EpiData and EpiData Analysis is freeware. This means that the program as such cannot be sold for money or service value. It is absolutely free. All translated versions must ALSO be supplied as free.

There can be NO charge taken by a web site for downloading of EpiData.

Introduction and Background

What is EpiData Analysis ?

EpiData Analysis is a program for data analysis and data management.

Use EpiData analysis when you want to do basic descriptive statistical analyses, modifications or tabulation of data. Extended analysis such as statistical modelling can be done with other software such as Stata, R etc.

EpiData Analysis is based on the same principles as EpiData Entry. If you properly define, document and verify data with EpiData Entry, the definitions are also available in EpiData Analysis. E.g. specified legal values with attached text labels(1 = No 2= Yes) or definitions of missing values. When reading data EpiData Analysis will do some control based on variable definitions, e.g. all dates are controlled.

EpiData Analysis is suitable for as well small as rather large datasets. simple datasets like one questionnaire as well as datasets with many or branching dataforms[1]. **EpiData** is freeware and available from Http://www.epidata.dk.

The principle of EpiData is rooted in the simplicity of the dos program Epi Info, which has many users around the world. The idea is that you write simple text lines and the program converts this to a dataentry form. Once the dataentry form is ready it is easy to define which data can be entered in the different data fields.

If you want to try EpiData analysis during the coming pages make sure you have downloaded the program and installed it.

It is an essential principle of EpiData not to interfere with the setup of your computer. EpiData consists of one program file and a few help files. No other files are installed. (In technical terms this means that EpiData does not install or include any DLL files or system files apart from screen fonts and a single file to create png graph files - options are saved in ini files)

Registration

All users are encouraged to registrate by using the form on www.epidata.dk . By registration you will receive information on updates and help us in deciding how to proceed development - and to persuade others to add funding for the development.

[1] Relate functionality has not been controlled fully in first releases of Analysis.

History.

During the period 2002-2003 Jens Lauritsen structured ideas of making an analysis programme to extend EpiData Data Entry possibilities and to implement an effective update of Epi Info v6 analysis in the windows environment. MD and epidemiologist Salah Mahmud joined the efforts and until summer/autumn of 2003 was in charge of programming design and implementation. Due to other commitments Salah Mahmud stopped working on the project. The intellectual value addition of Salah Mahmud is the core parser of EpiData Analysis.

Since autumn 2003 coordination and implementation has been done by JM. Lauritsen with Michael Bruus being in charge of procedures for reading data and chk files. A unified logical interface principle has been implemented, and a series of automatic test routines developed to certify results and estimates. The core estimation principles for tables is based on Epi Info version 6 source code. Several dedicated persons and users in general have been very active in testing and assistance of getting a uniform product developed.

The first public beta was released for testing in October 2004 with new builds released since periodically. During the development in the period until July 2005 the need for restructuring of the internal programming structures and unified logic has evolved and the implementation of an updated structure as shown on next page is ongoing. An updated version and status list is available as the document "revision.htm" installed with the beta version and published as http://www.epidata.dk/analysisinfo/docs/revision.htm.

Useful internet pages on Biostatistics, Epidemiology, Public Health, Epi Info etc.:

Data types and analysis:	http://www.sjsu.edu/faculty/gerstman/EpiInfo
Statistical routines:	http://www.oac.ucla.edu/training/stata/
Epidemiology Sources:	http://www.epibiostat.ucsf.edu/epidem/epidem.html
Epidemiology lectures:	http://www.pitt.edu/~super1/

Freeware for dataentry, calculations and diagrams:

EpiData (current program) for dataentry is available at www.epidata.dk

Epicalc 2000 Epidemiological oriented calculator. http://www.myatt.demon.co.uk/

EpiGram for drawing flowcharts and diagrams http://www.myatt.demon.co.uk/

OpenEpi Initiative:	http://www.cdc.gov/epo/epi/epiinfo.htm
Epi Info home page:	http://www.cdc.gov/epo/epi/epiinfo.htm

Steps in a documentation or research project

1 Aim and purpose of investigation is settled

- Hypothesis described, Size of investigation, time scale, Power calculation ...
- Funding ensured, Ethical commitee ....etc.

2 Ensuring Technical dataquality at entry of data

Collect data and ensure quality of data from a pure technical point of view. Document the process in files and error lists.

- done by applying legal values, range checks etc
- entering all or parts of data twice to track typing errors.
- finding the errors and correcting them

3 Consistent data and logical assertion.

The researcher cross examines the data. ”Get to know your data”. Trying to see if data are to be relied upon:

- Sound from a content point of view (no grandmothers below age of xx, say 35)
- Amount of missing data. Some variables might have to be dropped or part of the analysis should question influence on estimates in relation to missing.
- Decisions on number of respondents (N).

Describe the decisions in a document together with descriptions of the dataset, variable composition etc. In this part you would never do statistical testing.

4 Data Clean Up, derived variables and conversion to analysis ready dataset.

In most studies further clean-up and computation of derived variables is needed. E.g. in a follow-up study where periods of exposure should be established , merging of interview and register based information, computation of scales etc. Along this clean up decisions on particular variables, observations in relation to missing data are made. These decisions should all be documented.

5 Archive copy of data in a data archive or safety deposit. Include copies of all project plans, forms, questionnaires, error lists, other documentation. The aim is to be able to follow each value in each variable from final dataset to original observation.Archive original questionnaires and other paper materials as proof of existence in accordance with "Good Clinical Practice Guidelines", "Research Ethical Commitees" etc (e.g. for 10 years)

6 Actual analysis and estimation is done. All analysis is made in a reproducible way in principle. Sufficient documentation of this will be kept as research documentation.

EpiData Analysis is focused on steps 3. 4 and 6

Data documentation process

Aim and purpose of investigation is settled

- Hypothesis described, Size of investigation, time scale, Power calculation ...
- Funding ensured, Ethical commitee ....etc.

Ensuring Technical dataquality at entry of data

Collect data and ensure dataquality of data from a pure technical point of view

- done by applying legal values, range checks etc
- entering all or parts of data twice to track typing errors.
- finding the errors and correcting them

Documenting the process in files or error lists.

Consistent data and logical assertion.

The researcher cross examines the data. Trying to see if data are to be relied upon:

- Sound from a content point of view (no grandmothers below age of xx, say 35)
- Amount of missing data. Some variables might have to be dropped or part of the analysis should question influence on estimates in relation to missing.
- Decisions on number of respondents (N).

Describe the decisions in a document together with descriptions of the dataset, variable composition etc.

Data Clean Up, derived variables and conversion to analysis ready dataset.

In most studies further clean-up and computation of derived variables is needed. E.g. in a follow-up study where periods of exposure should be established , merging of interview and register based information, computation of scales etc.
Along this clean up decisions on particular variables, observations in relation to missing data are made. These decisions should all be documented.

Research Documentation

Archive copy of data in a data archive or safety deposit. Include copies of all project plans, forms, questionnaires, error lists, other documentation. The aim is to be able to follow each value in each variable from final dataset to original observation.
- Archive original questionnaires and other paper matereals as proof of existence in accordance with "Good Clinical Practice Guidelines", "Research Ethical Commitees" etc (e.g. for 10 years)

Technical principle: Memory and structures for handling of data:

This and the next two pages is rather techical in nature and can be skipped by beginners.

EpiData Analysis is based on the overall principle that users have good data, they want to preserve the good data as such and only modify data on disk by explicit commands. Commands are created by the experienced user directly in writing, by new users via menu's and the "Analysis Toolbar".

EpiData Analysis reads a copy of data into memory and only modify or handle the data in memory. If the file has an encrypted field analysis will ask for the pass key at reading of the data.

The assumption is that with structured commands the user can always replicate what was done by menu's, written commands or saved commands. Only by **specific** commands is information saved on disk. The commands for this are **savedata**, **savepgm** and **logopen** saving data files, executed commands and results. When the program closes down a copy of all commands is saved as temp.pgm with previous copies renamed as temp1.pgm, temp2.pgm and temp3.pgm.

Two structures control running of analysis. Set parameters and options. Set parameters are actually commands that specify certain features, e.g. the font size of output (**Set viewer font size=14**) whereas options give specific information to a given command, e.g. **freq a b c /nomissing** , which will select for the frequency table of a b and c only those records with information in all three variables.

In figure 1 next page the principle is shown in detail. Users can specify how the program is working by adding set commands to the startup file "Epidatastat.ini".

CHK file parsing into analysis.

The EpiData chk file language consists of more than 100 different commands and functions. These are documented in the help file of the EpiData Entry programme. EpiData Analysis will read **all datafiles (rec)** and associated chk files which can be read by the EpiData Entry module[2]. The following chk file structures are used actively in current analysis:

Structure	chk file command	example or comment
category labels associate explanatory text with numerical values in the data,	comment legal labelblock	comment legal 1 dog 2 giraffe end
Definition of zero to three numerical values indicative of missing or irrelevant data.	missing value	missing value 7 8 9 missing value 97 98 99

[2] Until fully tested some special chk commands might give problems in this relation.

The following structures are planned implemented for proper consistency checks of data, range, legal, mustenter and later possibly also jumps. Such that analysis can reproduce results of consistency checks now available by the EpiData Entry and EpiData Epic programmes.

Output formatting

EpiData Analysis output is based on the html (hyper text meta language) format. HTML is a structured formatting principle used on all internet pages. EpiData Analysis will comply with the W3C standards. W3C is an international organisation informing about developments and standards on Http://www.w3c.org.

Output formation is done in a combination of the programme creating structures and contents of the structures, whereas the way the structures are presented to the user is defined in a a simple text file (epiout.css). In this way it is easy for users to adapt the output format, e.g. to publish reports on a web site. Definitions contained in the simple text file is technically called a cascading style sheet (CSS) containing definitions for size, colour, margins etc. A full documentation of the output principles is available[3] The documentation is installed with the analysis programme.

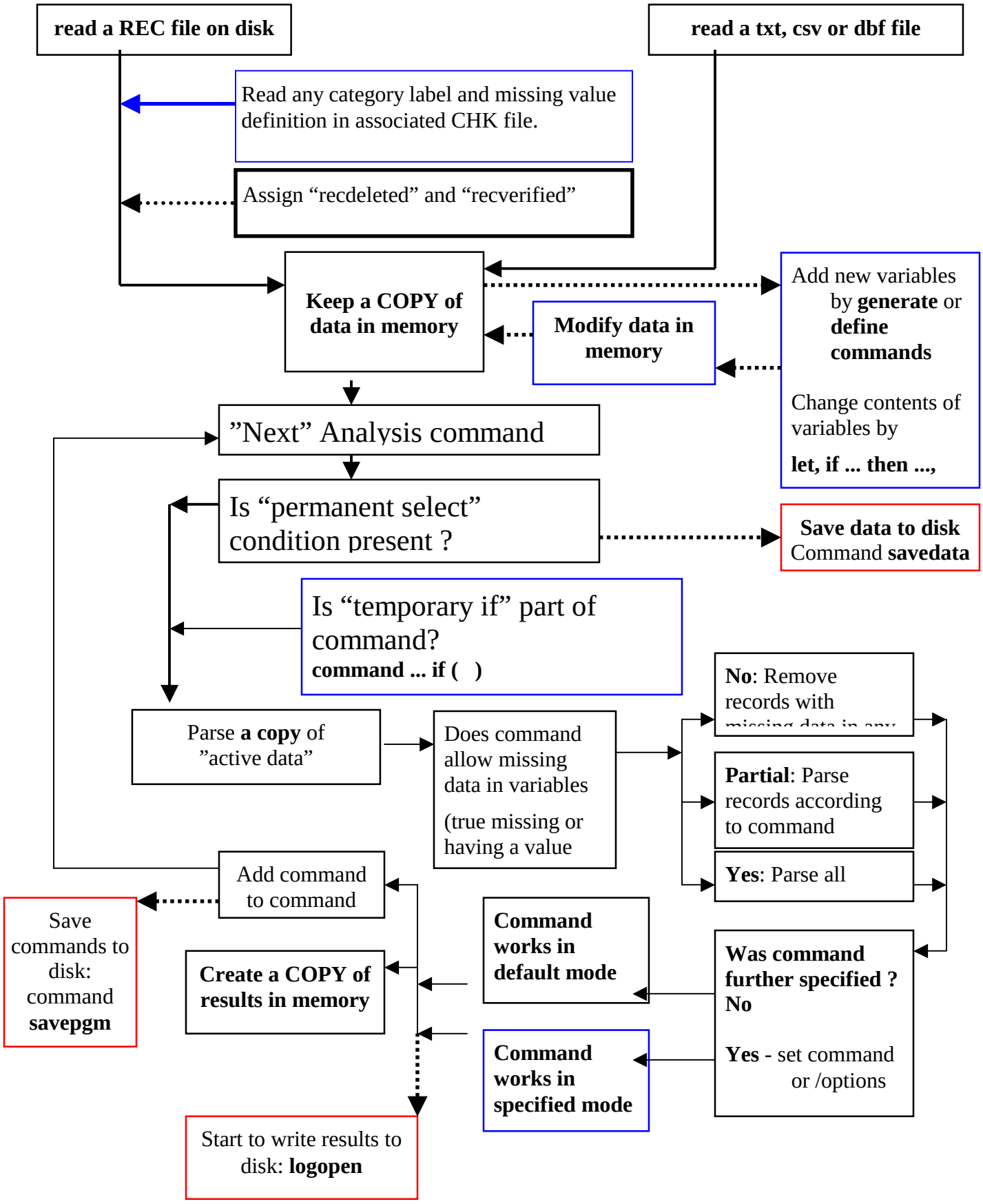
Flowsheet of how you work

Flowsheet of EpiData Analysis handling of data is shown on next page. (Next page)

The flow sheet might seem overly detailed – in particular for a new user, but when knowing how or when a certain aspect is applied refer to the flowsheet.

[3] Lauritsen JM. Output layout definition in EpiData Analysis v1.1. EpiData Association, Odense Denmark, 2005. Http:\\www.epidata.dk\analysisinfo\html output definition.htm

Black parts: In memory only. Blue parts depending on setup, specification or special options.
Red: save to disk



0. Install EpiData Analysis

Get the latest version from Http://www.epidata.dk and install in the language of your preference. The installation and retrieval is fast[4] since the whole size of the programme is small (2.5Mb in total). Simply run the installation file and follow the instructions.

1. First run and Setup screen of EpiData:

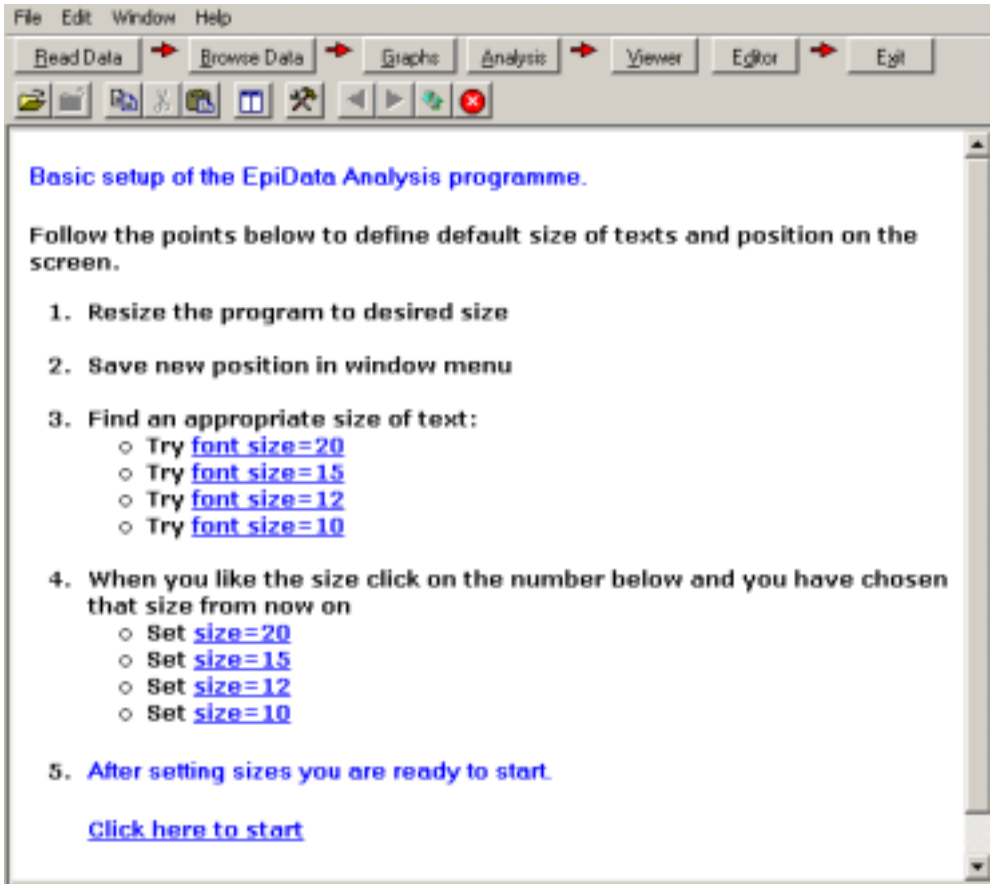
The EpiData Analysis screen has a "standard" windows layout with one menu line, a toolbar, an output window, a command prompt where you write commands, an output window, a statusbar and depending on user action also a data browser, an editor and smaller windows showing available commands, variables in current file and a history of previous commands.

To get acquainted with the programme only parts of this will be introduced in this document.

With experience you will get to know the other parts.

A. Start EpiData now from either the programme group where you installed the programme or an icon.

B. First time you run EpiData Analysis the opening screen will be a setup screen where you decide overall size of programme, size for fonts you prefer etc. Just follow the numbers:



C. After clicking start, the screen changes.

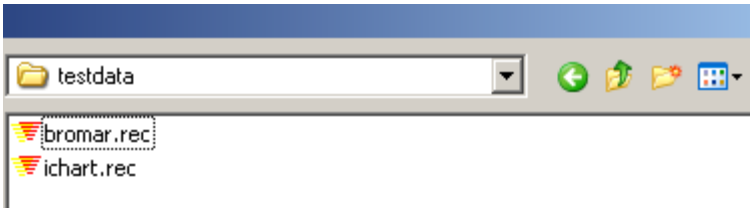
[4] If you are on a slow modem line you might not agree to "fast", but in comparison to many programmes this is a small size

2. First trial with programme:

To give you an idea of how the programme works we will now try the “Process Toolbar”:

Press the “Read Data” button to the left on your screen and point your disk selector to where you installed the EpiData Analysis programme in the subfolder “testdata”.

Select “bromar.rec” (possibly shown only as “bromar”). by double clicking and notice the information in the output window, How many records and fields ?



Data

The data are copies of completion time, age etc. for a Marathon

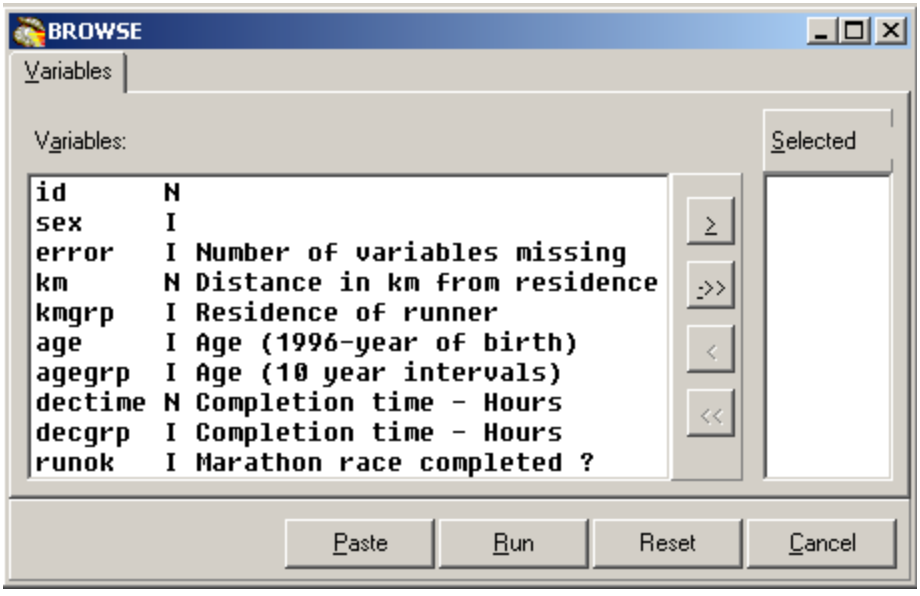
```
. read

Loading data C:\Programmer\EpiDatab36\testdata\bromar.rec, please wait..

File name : C:\Programmer\EpiDatab36\testdata\bromar.rec
Marathon data - 1995 across bridges from Funen and
Fields: 10 Total records: 4027 Valid records: 4027
```

Browse data

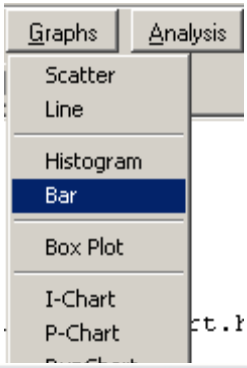
Every time we open a new data file it is good practice to view the data.

So on the process tool bar press the ”Browse” button, and you get this dialog box: Select any number of variables or all by the ->> and choose ”Run”.

A data browser window is shown. Look at the data. Resize the window to cover right half of the screen and close it again.



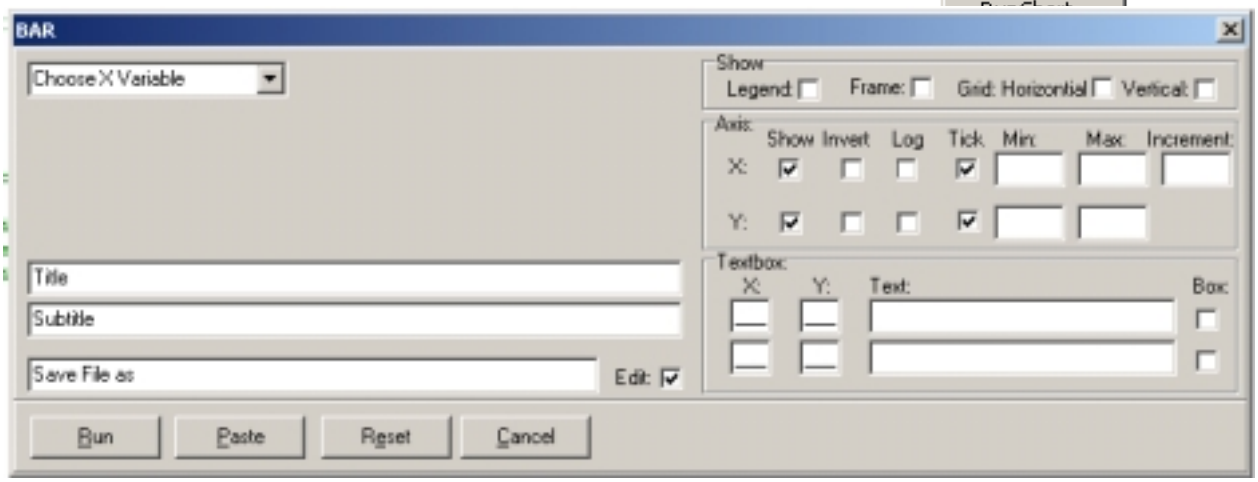
Showing a graph



Now we will try to see some results. Press the **"Graph"** button and you see the small menu, where you choose **"Bar"**

A dialog is shown to specify the graph further:

At this point only worry about the "**Choose X Variable"**. Select **Decgrp** and press **"Run"**

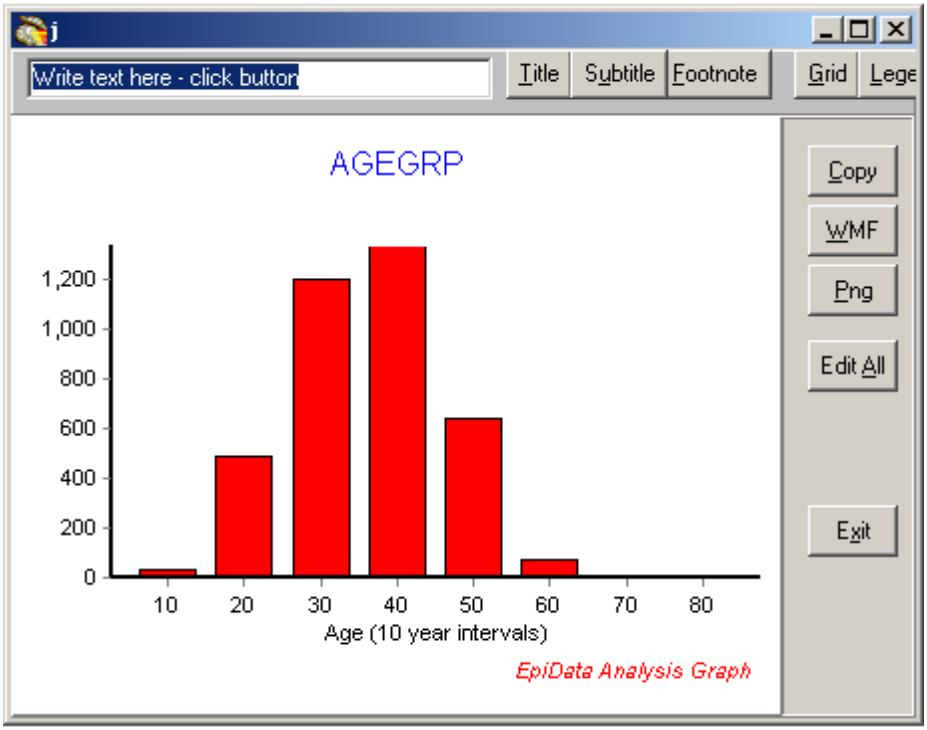


The graph is shown on a special form with some buttons.

Just press "exit"

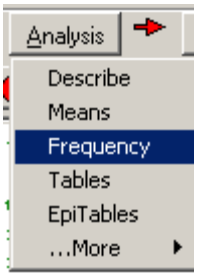
The graph will be added to the output window.

Try with kmgrp or decgrp variables and notice how the labels are shown at the buttom of the graph..



Showing a table of frequencies

On the ”process toolbar” try now analysis:



Select sex and agegrp variables in the dialog which is shown and press ”run”, after which you will see these tables:

```
. FREQ agegrp  sex
```

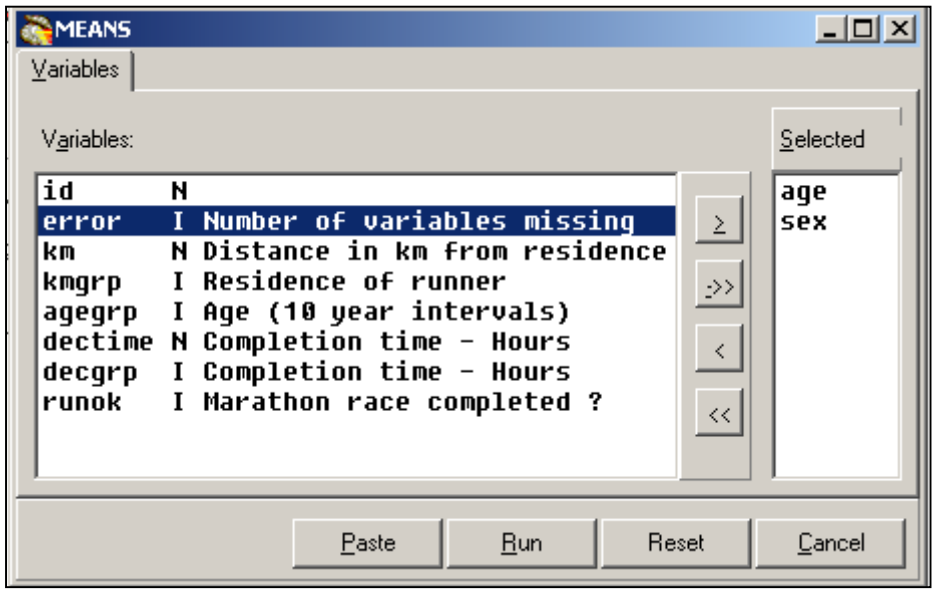
AGEGRP	No.	%
10	35	0.92
20	489	12.92
30	1198	31.64
40	1337	35.31
50	642	16.96
60	75	1.98
70	9	0.24
80	1	0.03
Total	3786	100%

SEX	No.	%
Female	490	12.17
Male	3537	87.83
Total	4027	100%

Showing means of age by time

Choose in Analysis instead ”Means” and ”run”

Are males or females on average the youngest ?



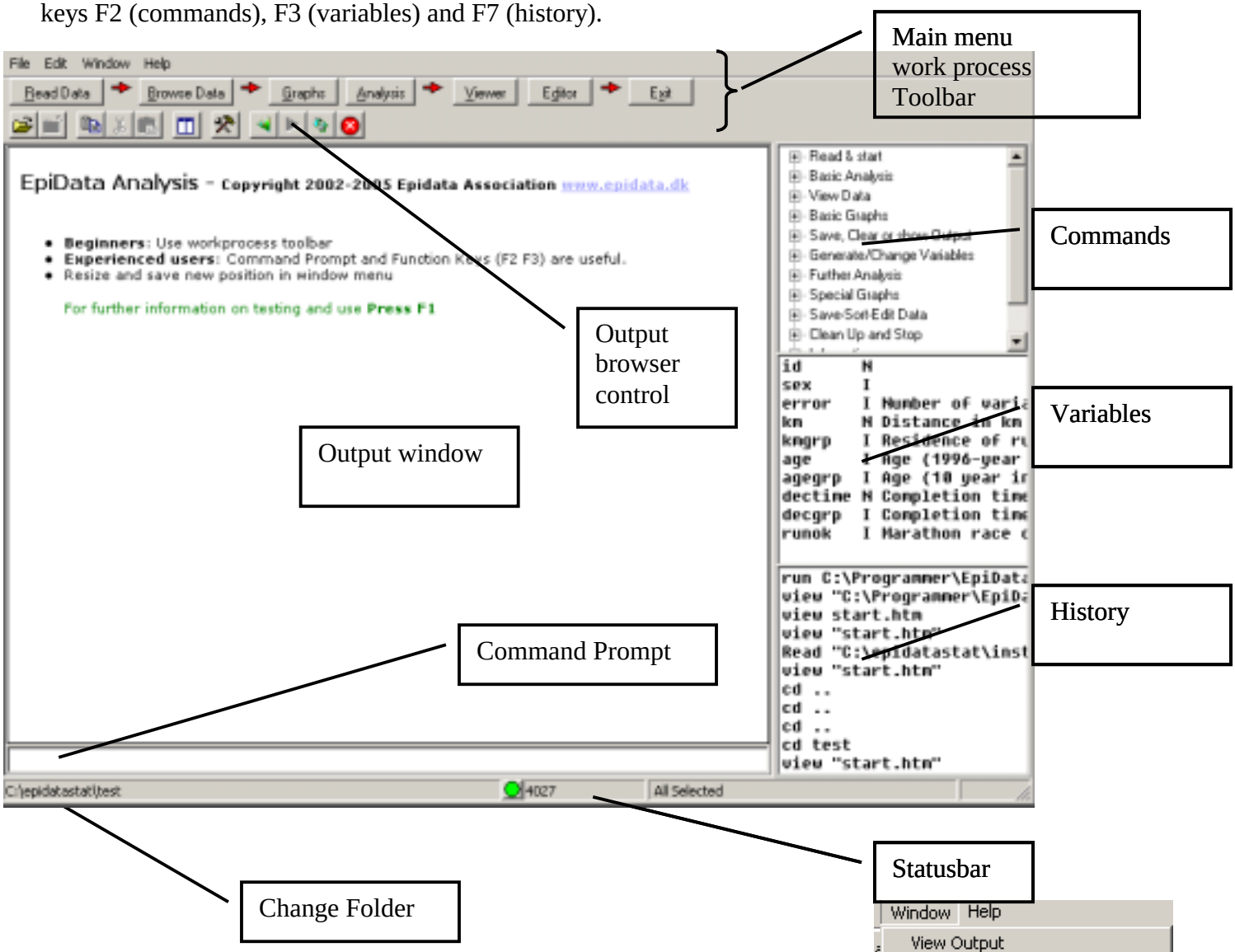
Age (1996-year of birth)

SEX	Obs.	Sum	Mean	Variance	Std Dev	(95% CI	mean)	Std Err	
Female	463	20094.0	43.40	77.97	8.83	42.59	44.21	0.41	
Male	3323	133646.0	40.22	97.60	9.88	39.88	40.55	0.17	
SEX	Minimum	p5	p10	p25	Median	p75	p90	p95	Max
Female	19.00	28.00	30.00	38.00	44.00	50.00	53.00	56.00	70.00
Male	16.00	24.00	27.00	33.00	40.00	47.00	53.00	56.00	84.00

3. Which elements are on the screen ?

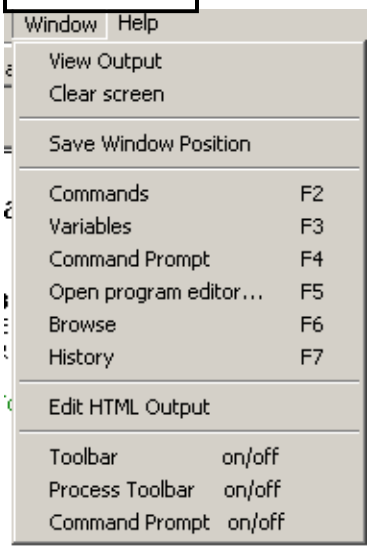
You have now acquainted your self with the programme and working with the **process toolbar.** In the next sections more aspects are covered with a bit more detail.

To see a picture like this: Start EpiData Analysis, read a datafile into memory and press function keys F2 (commands), F3 (variables) and F7 (history).



To get acquainted with the windows try the following:

1. Switch the extra windows on the right on and off a few times: Press keys: F2, F3 or F7
2. Resize the program by dragging in sides or the separator between output window (viewer) and right side parts.
3. Save current position in window menu. ”Save Window Position”
4. Try to change folder by clicking on the lower left side of the statusbar.



Further introduction will be added here

- how to create new variables. Raw and grouping
- how to save and print output
- how to append several files

Support

If you find errors or bugs when using the program or have suggestions for improval please discuss with the EpiData-list available at http://lists.umanitoba.ca/mailman/listinfo/epidata-list

Sources for support:
1..Read the help file to epidata.
2..Read this epitour document
3..Download from http://www.epidata.dk the epidata help file
and the epitour help file in the format of "pdf", which is easy to print.
4. Basic aspects of epidata follows the Epi Info version 6 manuals. This is available from the Epi Info site: http://www.cdc.gov/Epi Info/

Unfortunately we do not have resources for support of dataentry questions in general refer these to the EpiData-list available at http://lists.umanitoba.ca/mailman/listinfo/epidata-list

Funding and acknowledgements.
An updated list of attained funding is available at Http://www.epidata.dk/funding.htm and further credits and acknowledgements at: Http://www.epidata.dk/credit.htm . International translations made to several languages, see Http://www.epidata.dk
For donations to further development see help file or send an e-mail to info@epidata.dk

Disclaimer