



v1.0 2008

EpiData Introduction Guide – A Canadian Example

Created as part of a collaborative project¹ :
Association of Public Health Epidemiologists in Ontario (APHEO),
EpiData Association and the Public Health Agency of Canada.

Authors: APHEO EpiData Expert Panel and J.Lauritsen, EpiData Association.

Notice: This is a preliminary version of a field guide as part of the project. The field guide is “in development”. Final version to be released at a later date. Order of sections will most likely change.

Purpose and scope:

This introductory text is created for self-instruction or a short facilitated seminar of 3 to 4 hours in length. Actual length depends on the number of facilitators and the audience’s experience with EpiData.

Participants should have prior understanding of outbreak investigation methods and the associated statistical methods.

The structure of this guide includes the principles of defining and entering data followed by basic analysis principles for a simplified dataset, and finally in part three, follow-up exercises for the experienced user or a follow-up seminar. Part three would normally not be part of the first 3 to 4 hour seminar. Data used herein are partially modified for exercise purposes and cannot be published except by written permission.

This instruction is based on the study “Bridal Shower Outbreak” By York Region Community and Health Services, Ontario, Canada

Introduction - Background of Outbreak

An outbreak of gastrointestinal illness was declared by York Region Community and Health Services on February 26, 2008, following a bridal shower held at a banquet hall on February 24, 2008. Approximately 75 guests were served meals, and 36 employees worked as food handlers as part of this event. Food history questionnaires were administered to 48 symptomatic and 16 asymptomatic persons. EpiData Entry was used to create the questionnaire form and the data was then entered.

A retrospective cohort design was used to collect and analyze the data. A retrospective cohort study relies on historical exposure data. In this outbreak, a representative sample of participants was interviewed, allowing the comparison of the rate of disease in those who did or did not eat a certain food. This type of study is the technique of choice when one is faced with an acute outbreak in a well-defined population.

¹ This project has been made possible through a financial contribution from the Public Health Agency of Canada

Introduction – EpiData Software

EpiData is a free software suite designed to assist epidemiologists, public health investigators and others to enter, manage and analyze data in the field. All software is available from <http://www.epidata.dk>. A number of field guides, software documentation notes, examples and other information are also available. Users are encouraged to:

- (1) Join the EpiData-list discussion group, and
- (2) Sign up for the information newsletters sent periodically each year. For more information, visit the EpiData website at <http://www.epidata.dk>

Preparation

Do the three steps A, B and C:

A. Read about data management and structures

Go through the document:

http://www.epidata.org/wiki/index.php/Field_guide:_from_question_to_data which discusses data structures, data management and documentation.

B. Acquire and install software

Download EpiData Analysis (v2.1 or later) and EpiData Entry (v3.1 or later) and install these on your computer. You should use the most recent version available from the download part of the website <http://www.epidata.dk>. In case installation on your computer is blocked by an administrator, you can install into your private folder (create a subfolder on the desktop) or resort to installation on a USB key².

NOTE: October 2008 – if Analysis version 2.1 is not on the general download page <http://www.epidata.dk/download.php>, then find it on <http://www.epidata.dk/testing.php>

C. Starting a New Project

Create and name a new folder

You can create a folder for your investigation anywhere you like on your computer; you do not have to create one in the same drive as the EpiData program files.

Get the necessary exercise data file for this note from the seminar facilitators and UNZIP this file into your project folder. If you have no other unzip software, EpiData Entry has a utility in the main menu.

² But it is good practice to talk with the administrator also. All EpiData Software is checked with up-to date anti-virus and other malicious software protection agents before uploading to [epidata.dk](http://www.epidata.dk).

PART 1: EpiData Entry

Data entry

Content to follow

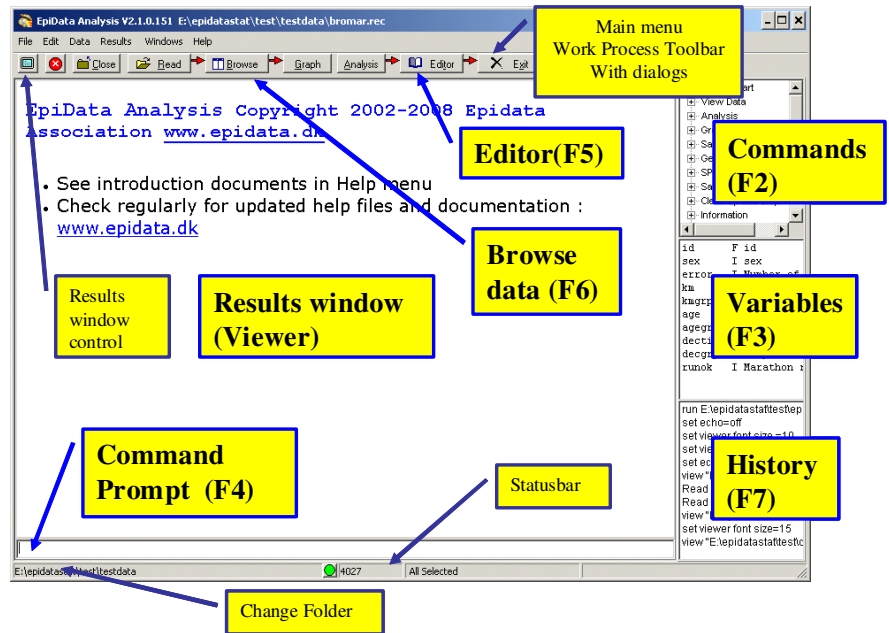
PART 2: EpiData Analysis

EpiData Analysis is used to analyze EpiData *.rec files, dBase *.dbf files, and text *.csv files.

The EpiData Analysis screen is easy to navigate and functionality is accessible via the use of shortcut keys.

For the purposes of this field guide, areas of importance are:

- The command prompt (F4) area located at the bottom of the screen
- The editor (F5)
- The results window.
- The dialogs in the work process toolbar.



But also take some time to explore additional functionality available via other short cut keys.

When you open EpiData Analysis, you see the following menu:

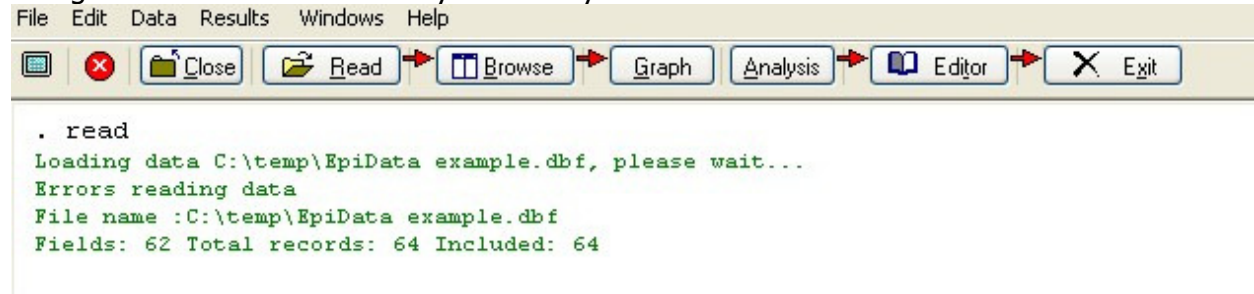


Opening and preparing the dataset

To open a data file, use:

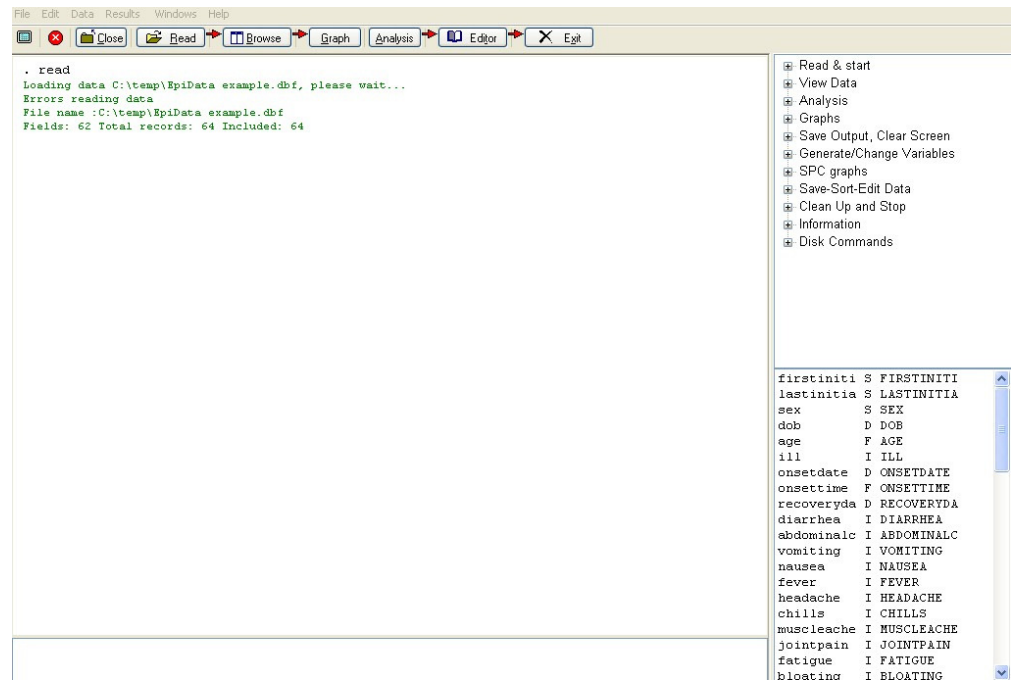


Navigate to the location where you saved your file and select it.



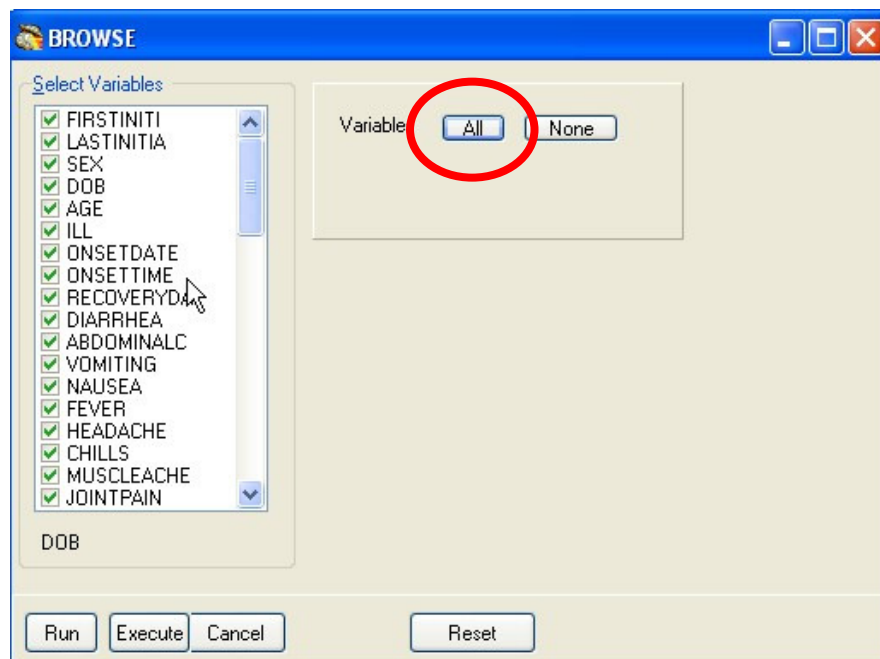
EpiData will provide some information about the data: name, number of records number of fields, etc. In this case we have 64 records in the dataset.

Press F2 and F3 to reveal all commands as well as the file's variables:



Now take a look at the data: Use:





To get a line listing:

	FIRSTINITI	LASTINITIA	SEX	DOB	AGE	ILL	ONSETDATE
1		M	F	31/12/1958	49	1	25/02/2008
2	P	C	F	14/02/1989	19	1	25/02/2008
3	L	A	F	05/07/1971	36	1	24/02/2008
4	R	L	F	08/09/1945	62	1	25/02/2008
5	L	F	F	18/08/1962	45	1	25/02/2008
6	M	P	F	07/07/1954	53	1	25/02/2008
7	R	M	F	21/10/1998	9	1	26/02/2008
8	G	C	F	27/02/1972	36	1	26/02/2008
9	G	E	F	10/08/1982	25	1	25/02/2008
10	G	M	F	10/10/1956	51	1	25/02/2008
11	G	N	F	26/03/1975	32	1	25/02/2008
12	G	A	F	16/11/1999	8	1	25/02/2008
13	V	L	F	23/11/1999	8	1	26/02/2008
14	V	E	F	27/01/1970	38	1	26/02/2008
15	M	D	F	26/08/1967	40	1	26/02/2008
16	D	A	F	07/02/1957	51	1	25/02/2008
17	G	L	F	.	48	1	25/02/2008

Fine tuning the dataset for analysis (data cleaning)

Before analysis can be done in a valid way, the data must be cleaned and documented – if it hasn't already been done during entry. This includes finding out whether all variables are valid, how many observations can be part of the analysis etc. Here a few examples are shown; more aspects will be covered in the extended part 3 of this note.

1. Changing Variable and Value Labels

Sometimes when you start analyzing your dataset, you realize that the names of variables or values are not all that meaningful. In particular in these instances, it is important to spend some time preparing the dataset, but it is always good practice to define labels at three levels:

- (1) At whole file level (labeldata),
- (2) At variable level (label), and
- (3) At value or category level ((labeldata))

- To add (or change) variable names, use the command: LABEL
In the command prompt area (F4) write: *LABEL LEMONSORBE "lemon sorbet"*
- To add labels to the values of LEMONSORBE use the command: LABELVALUE
In the command prompt area (F4) write:
LABELVALUE LEMONSORBE /0="No" /1="Yes" /2="Unknown"
- To add labels to the whole data file: LABELDATA
In the command prompt area (F4) write:
LABELDATA "Bridal shower outbreak investigation"

There is an easy way to add value labels for many variables in sequential order:

```
labelvalue nonalcohol-friedfishc /0="No" /1="Yes" /2="Unknown"
```

* Notice: no space btw the variable names only one dash: -

Repetition Question:

What is a data label, a variable label and a value label used for and what commands are used for this?

2. Creating New Variables and Adding Conditional Values

During the data entry process, the age of each case and non-case was collected in two ways. Each was asked the birth date, and also the age at investigation. Now that we want to analyze the data, we want to group ages into categories. For now we will only use the recorded age, whereas calculation of age based on Date of Investigation and Date of Birth follows in part 3.

We are going to use the "DEFINE" and "RECODE" commands.

In the command prompt (F4) area, write:

* create your age group variable as an integer:

```
define agegrp #
recode age to agegrp 0-9=1 10-19=2 20-44=3 45-64=4 65-hi=5
```

The command "recode" also adds the value label in the intervals indicated

Repetition Question:

What is the difference between integer, float, string and date variables?

Descriptive Epidemiology

Describing the cases in terms of person, place and time is good epidemiological practice and can help you develop hypotheses about the mode of transmission and the source of infection. Let's start with **person**:

In order to know how many cases and non-cases we have we can do a frequency distribution of the variable ILL.

In the command prompt area (F4), write:

```
FREQ ILL
```

You can see that there are 48 cases and 16 non-cases.

* (Where 0=non-case and 1=case)

Now, analyze the demographic data to determine the age and sex structure of the population at risk. Use the new variable you just created above – agegrp. Once again, at the command prompt (F4), type:

```
tables sex agegrp /c
```

This will produce a cross tabulation of sex by age group along with column percents. Your table should look like this:

agegrp	SEX		Total	%
	F	% M		
0-9	7 {13.0}	2 {20.0}	9	{14.1}
10-19	2 {3.7}	2 {20.0}	4	{6.3}
20-44	20 {37.0}	4 {40.0}	24	{37.5}
45-64	21 {38.9}	1 {10.0}	22	{34.4}
65+	4 {7.4}	1 {10.0}	5	{7.8}
Total	54 {100.0}	10 {100.0}	64	

Percents: {Col}

You can see that the majority of the population is between the ages of 20 and 64 years and that more than 80% are women. We can add row percents as well by adding /r to the end of the command line.

Use:

Analysis

To, e.g. DESCRIBE the content of a continuous numeric field such as AGE and get:

Variable	N=64	Sum	Mean	(95% cfi)	Min	p5	p10	p25	Median	p75	p90	p95	Max
AGE	64	2353.00	36.77	31.77 41.76	0.0	0.250	8.50	21.25	37.00	51.75	62.00	69.00	78.00

For the variable AGE, the total number of records used in the calculation are given (N=64) as well as the sum, mean age, 95% Confidence Interval, the minimum value, percentiles (5, 10, 25, 50=Median, 75, 90 and 95) and the maximum value.

Creating an Epidemic Curve

Describing your cases with respect to **time** will give you clues to understanding the mode of transmission of the disease, the source and the nature of the etiologic agent (incubation period).

An epidemic curve is a graph that provides a visual display of the magnitude and time course of an outbreak. The epidemic curve plots time along the X-axis and number of cases along the Y-axis. Because time is continuous, the epidemic curve is drawn as a histogram (no gaps between adjacent columns).

Before creating the epicurve, select only those people that were ill – the non-cases are not contributing to analysis of date of onset.

In the command prompt (F4), write:

Select ill=1

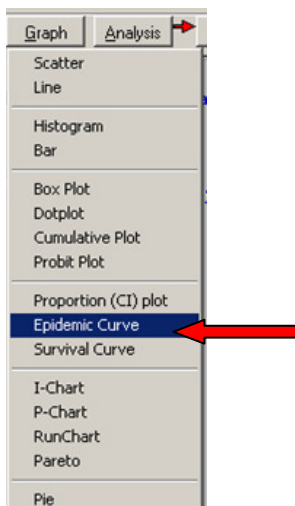
Notice how the status bar at the bottom changes with the select command.

Question: In what way is *select* reflected in the status bar?

Use:

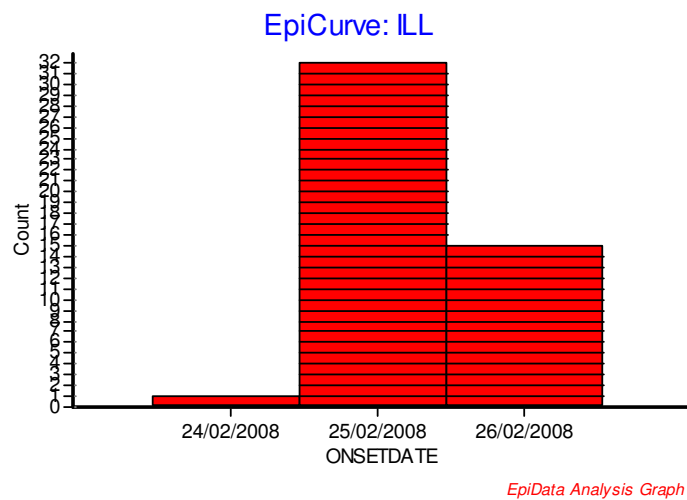
Graph

to make a variety of graphs, including the Epidemic Curve



Now, do the epidemic curve based on the graph dialog for epidemic curves, which will formulate the command as:
Epicurve ill onsetdate

The graph's default title and look:



Question: What would happen if we forgot to select cases?

Try: You revert to all cases by writing the command 'select' with no statement after in the F4 area:

Select

Redo the Epidemic curve and see if there is a difference to the first one.

Hint: You may scroll through previous commands in the F4 area with the Up-Arrow.

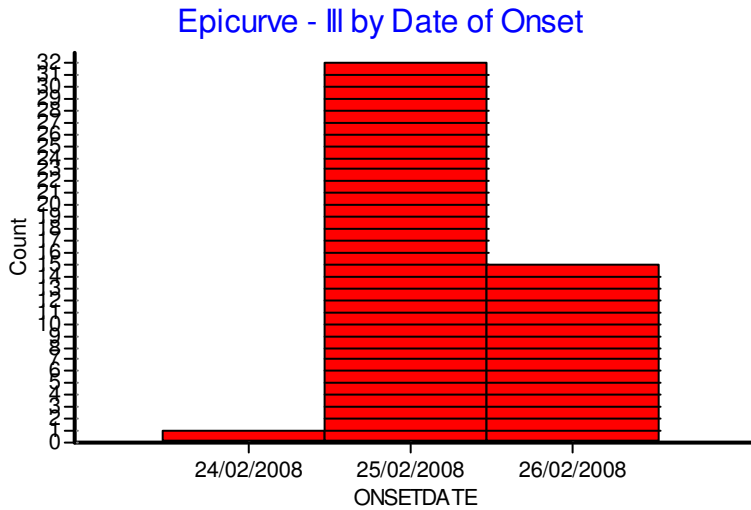
Answer: You will notice there is no difference, since the command *Epicurve* apparently bases the information only on cases. – But now you know a way to analyze a subgroup with *select*

Options:

Move on with the next exercise to work a bit with options. Options are specifications which make a command behave according to further details given by the user. Many commands have options, which are documented in the help file system, which is invoked by the F1 button.

Let's change the title by writing in F4:

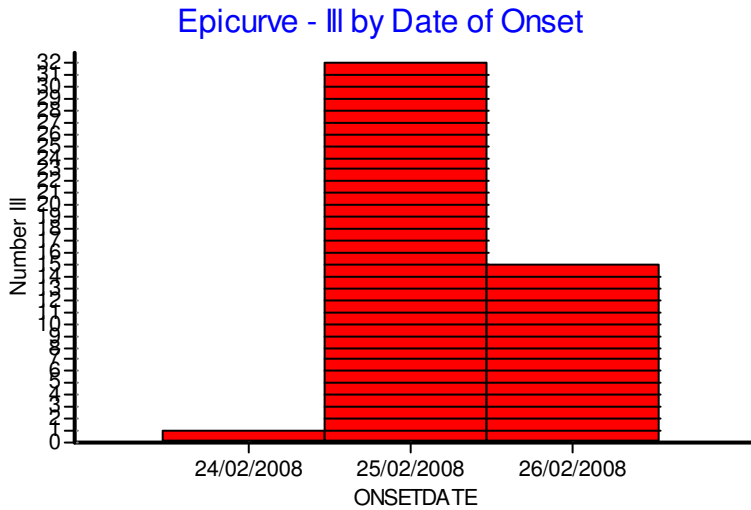
Epicurve ill onsetdate /ti="Epicurve – Ill by Date of Onset"



EpiData Analysis Graph

Now, change the y axis label:

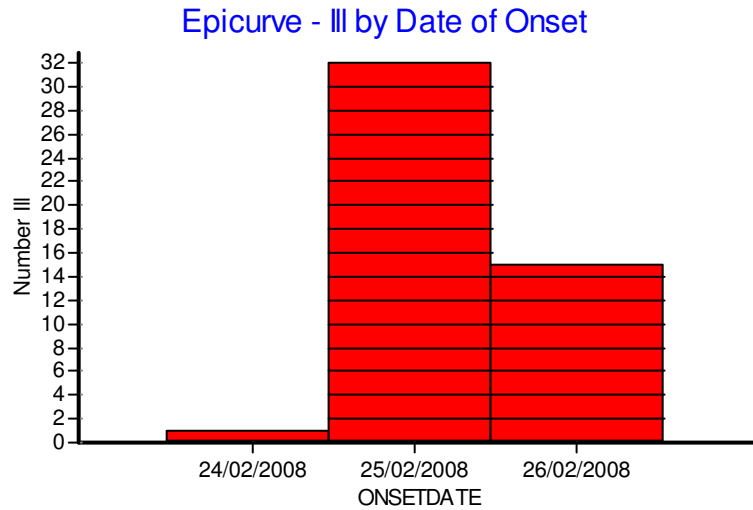
Epicurve ill onsetdate /ti="Epicurve - Ill by Date of Onset" /ytext="Number Ill"



EpiData Analysis Graph

The y-axis numbers are tight. Change the increment to 2:

Epicurve ill onsetdate /ti="Epicurve - Ill by Date of Onset" /ytext="Number Ill" /yinc=2



EpiData Analysis Graph

Now, the graph is much easier to interpret. The epicurve shows that the onset date of symptoms predominantly occurred during a 48-hour period between February 25 and February 26, 2008, approximately 24 hours after the event took place. The epidemiologic curve indicates a point source outbreak with no observed secondary transmission of infection.

Analytical Epidemiology

Food Specific Attack Rate Table

At this point in our retrospective cohort study, we are hoping to identify risk factors that might indicate the cause and mode of transmission of the disease. We created a questionnaire that asked both persons who were exposed and those who were not exposed to different foods and beverages if they became ill. What we are trying to do is determine the probability that a food item (for example, eating lemon sorbet) is linked to the outcome (becoming ill) by calculating attack rates for each food item.

The food item that is the true source of infection will likely have three features:

- a) The attack rate is high among persons who ate the food (high food-specific attack rate).
- b) The attack rate is low among persons who did not eat the food (so the difference or ratio is high).
- c) Most of the cases were exposed, so the exposure could “explain” most, if not all, of the cases.

Before we begin, remember: your select statement from the epicurve analysis is still valid if you did not turn it off as indicated above. If you want to create an attack rate table of

ill/not ill by food consumed, you must remove your selection (ill=1). This is easy to do. In your command prompt (F4), type 'select'. Verify that the command is correct by typing:
Freq ill

Both ill/not ill should appear:

ILL	
	N
0	16
1	48
Total	64

Attack rate tables in EpiData Analysis are possible by using the 'tables' command.

The general syntax for TABLES is:

TAB OUTCOME EXP1 [EXP2...EXP#] /OA [other options]

where:

OUTCOME is the variable where you classify your cases (default value for case is the highest in that variable). Use /SD to reverse that order. EXP# is a list of exposure variables. /OA indicates the option to create an outbreak table

In the command prompt window (F4), type the following command:

tables ill lemonsorbe /ct /ar

where ct = compact tables, and ar = attack rates. /oa is short form for "/ct /ar"

The following output appears:

```
. tables ill lemonsorbe /ct /ar
```

Outcome: ILL by LEMONSORBE									
ILL by LEMONSORBE	N	Exposed			Not Exposed			RR	(95% CI)
		n	Ill	AR (%)	n	Ill	AR (%)		
	61	44	42	{95.5}	17	4	{23.5}	4.06	(1.72-9.58)

Exposure: (LEMONSORBE = 1)

Outcome: ILL = 1

The results indicate that for those that ate (were exposed to) the lemon sorbet (n=44), 42 became ill for an attack rate (AR) of 95.5%. For those that did not eat the sorbet but were ill, the attack rate was 24%. The relative risk (RR) is also significant, although the 95% confidence interval is quite wide. The RR implies a strong association between the consumption of lemon sorbet and becoming ill.

Repeat the command typing in all of the food items:

tables ill lemonsorbe prosciutto caprese grilledveg cantalope fruttadima pasta farfelle freshtomat cannelloni freshherb vealmarsal chickenbre stirfryveg ovenroaste gardensala freshfruit coffee tea espresso pop juice mineralwat springwate redwine whitewine cheesecake chocolatec othercake shrimppatt pastries cookies tarts friedfishc /ct /ar

Your results will look like this (only partial table shown):

`tables ill grilledveg cantalope pasta farfelle freshtomat cannelloni`

Outcome: ILL

	ILL	N	Exposed		Not Exposed		RR	(95% CI)
			n	Ill	n	Ill		
GRILLEDVEG	61	47	38	{80.9}	14	7	{50.0}	1.62 (0.94-2.78)
CANTALOPE	62	33	30	{90.9}	29	16	{55.2}	1.65 (1.17-2.33)
PASTA	45	35	30	{85.7}	10	3	{30.0}	2.86 (1.10-7.44)
FARFELLE	57	39	33	{84.6}	18	8	{44.4}	1.90 (1.12-3.25)
FRESHTOMAT	57	36	31	{86.1}	21	10	{47.6}	1.81 (1.13-2.89)
CANNELLONI	62	47	42	{89.4}	15	4	{26.7}	3.35 (1.44-7.80)
FRESHHERB	57	41	37	{90.2}	16	5	{31.3}	2.89 (1.39-6.01)
VEALMARSAL	60	31	24	{77.4}	29	21	{72.4}	1.07 (0.80-1.43)
CHICKENBRE	60	36	29	{80.6}	24	16	{66.7}	1.21 (0.87-1.67)
STIRFRYVEG	59	40	33	{82.5}	19	12	{63.2}	1.31 (0.90-1.89)
OVENROASTE	61	41	34	{82.9}	20	12	{60.0}	1.38 (0.94-2.03)
GARDENSALA	61	41	34	{82.9}	20	12	{60.0}	1.38 (0.94-2.03)

Exposure: (GRILLEDVEG = 1)
 (CANTALOPE = 1) (PASTA = 1) (FARFELLE = 1) (FRESHTOMAT = 1)
 (CANNELLONI = 1) (FRESHHERB = 1) (VEALMARSAL = 1) (CHICKENBRE = 1) (STIRFRYVEG = 1)
 (OVENROASTE = 1) (GARDENSALA = 1)
 Outcome: ILL = 1

You can also add a test of statistical significance to your output. Using the ill/lemon sorbet as an example, type the following:

tables ill lemonsorbe /ct /ar /T

where ct = compact tables, ar = attack rates and T = Chi square

The following output appears:

```
. tables ill lemonsorbe /oa /t
```

Outcome: ILL by LEMONSORBE											
ILL by LEMONSORBE	N	Exposed		AR (%)	Not Exposed		AR (%)	RR	(95% CI)	Chi ² (df=1)	p
		n	Ill		n	Ill					
	61	44	42	{95.5}	17	4	{23.5}	4.06	(1.72-9.58)	34.209*	0.0000

Exposure: (LEMONSORBE = 1)

Outcome: ILL = 1

*: Small Expected Numbers, use P_{exact} /ex ←

The warning message indicates that an expected value of less than 5 occurred in one or more cells. Therefore, the Fisher exact P value needs to be used rather than Chi square. Change the command to the following:

```
tables ill lemonsorbe /ct /ar /ex
```

where ct = compact tables, ar = attack rates and ex = Exact test

The following output appears:

Outcome: ILL by LEMONSORBE											
ILL by LEMONSORBE	N	Exposed		AR (%)	Not Exposed		AR (%)	RR	(95% CI)	P _{exact}	
		n	Ill		n	Ill					
	61	44	42	{95.5}	17	4	{23.5}	4.06	(1.72-9.58)	0.0000	

Exposure: (LEMONSORBE = 1)

Outcome: ILL = 1

The p-value is extremely small, suggesting statistical significance and a strong association between the consumption of lemon sorbet and becoming ill.

Stratified Analysis using CIplot

Stratified analysis involves examining the exposure-disease association within different categories of a third factor. It is an effective method for looking at the effects of two different exposures on disease. Looking at the food attack rate table above shows that more than one exposure (food item) had elevated relative risks and statistically significant p values.

A plot of proportions of cases for subgroups defined by other variables can be an effective way of finding high or low risk groups quite quickly. Define a search pattern for subgroups before you do any analysis and be careful with statistical testing, since a broad search and test strategy will modify any p-value you should regard as significant.

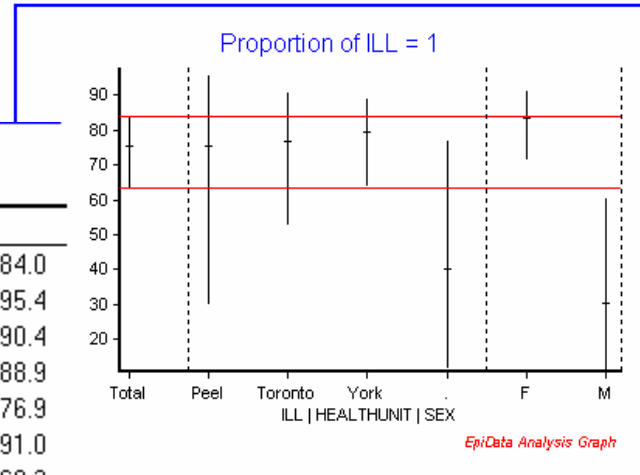
We want in this example to look for variation of illness in subgroups defined by other variables. Let us use sex and residence:

```
. cplot ill healthunit sex
```

Crude: Proportion of ILL = 1 among all.

variable	stratum	Total N	n _{ILL=1}	%	(95% CI)
ILL	Total	64	48	75.0	63.2 84.0
HEALTHUNIT	Peel	4	3	75.0	30.1 95.4
	Toronto	17	13	76.5	52.7 90.4
	York	38	30	78.9	63.7 88.9
	.	5	2	40.0	11.8 76.9
SEX	F	54	45	83.3	71.3 91.0
	M	10	3	30.0	10.8 60.3

Crude: Proportion of ILL = 1 among all.



Cplot ill agegrp sex healthunit

We notice that the proportion of ill persons is rather low among males and those with in Peel Region Health Unit and those with an unknown health unit (= period). The numbers are small, but an immediate suspicion could be that the low proportions in the two factors are the same male persons.

This could be investigated by further crosstables:
tables sex healthunit ill /M where M=show missing data

```
. tables sex healthunit ill /M
```

SEX			
HEALTHUNIT	F	M	Total
Peel	4	0	4
Toronto	13	4	17
York	33	5	38
.	4	1	5
Total	54	10	64

Unstratified table

ILL: 0 SEX			
HEALTHUNIT	F	M	Total
Peel	1	0	1
Toronto	2	2	4
York	4	4	8
.	2	1	3
Total	9	7	16

ILL: 1 SEX			
HEALTHUNIT	F	M	Total
Peel	3	0	3
Toronto	11	2	13
York	29	1	30
.	2	0	2
Total	45	3	48

Repeat the CIPILOT command (called "Proportion (CI) Plot" in the graph dialog part of the Work Process Toolbar) with the agegrp variable you created earlier. Is there any sign of variation in the proportion of illness by age?

Part 3: Working with programs

When you are analyzing a dataset it's good practice to save your commands so you can use them later and you don't have to 'recreate the wheel'. This is especially useful if you are working with a database in a routine system like a surveillance system.

You should use programs if you want to recode variables, make calculations using other variables or any other kind of manipulation. This method does not make permanent changes to your original dataset.

Saving, recalling and executing programs:

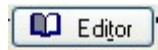
One way to make a program file is to enter commands interactively at the command prompt (F4). If you wish to save the commands you've already entered, use the SAVEPGM command and the name of a file, for example:

SAVEPGM OUTBREAK.PGM

All the commands used during the working session will be placed in the file, which can then be edited later to remove unwanted commands or add new ones.

You can also use the EpiData Analysis Editor.

Use:



And insert the commands with the menu part of the editor "insert history".
Once you have saved your commands in a PGM file, you can open it whenever you want using the EpiData Editor (FILE-->Open)

To execute a program you have different alternatives:

- i) In the EpiData Editor you can run all (F9)
- ii) Or in the EpiData Editor you can select the group of lines of the program that you want executed and then run the selected lines only (F8). With no selection only the current line is executed.
- iii) Edit the program in the editor, save, and from the command prompt issue the "RUN" command and find the .pgm file you saved.

Part 4: Extensions to the exercises

-
- Creating Age category definitions.

There are two approaches:

A. We are going to use the "DEFINE" and "IF [logical condition] THEN [action]" commands.
In the command prompt (F4) area, write:

```
define agegrp # (create your age group variable as an integer)
if age >= 0 and age < 10 then agegrp = 1
if age > 9 and age < 20 then agegrp = 2
if age > 19 and age < 45 then agegrp = 3
if age > 44 and age < 65 then agegrp = 4
if age > 65 and age < 79 then agegrp = 5
* if some people have missing age, they will not get a value
* with the above statements, therefore we give the value 9:
if agegrp = . then agegrp = 9
```

* Assign value labels:

```
labelvalue agegrp /1="0-9" /2="10-19" /3="20-44" /4="45-64" /5="65+"
```

B. The one used in part 1.

We are going to use the "DEFINE" and "RECODE" commands.

In the command prompt (F4) area, write:

* create your age group variable as an integer:

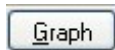
```
define agegrp #
```

```
recode age to agegrp 0-9=1 10-19=2 20-44=3 45-64=4 65-hi=5
```

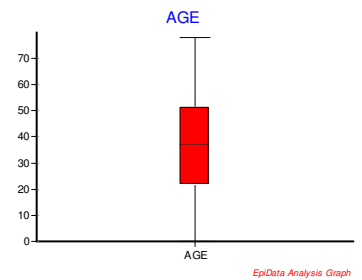
The command "recode" also adds the value label in the intervals indicated

- It is always good practice to view data in a graph which, for age, can conveniently be done with a box-plot and/or a cumulative plot:

Use:



And draw a box plot of age, a cumulative plot and a probit plot.



Q: Does the shape of the probit plot indicate that the age variable can be analyzed assuming Gaussian distribution?