Thực hành dịch tễ học thú y

Ths. Lê Thanh Hiền

Mục Tiêu

- **Giới thiệu** chung một số phần mềm sử dụng trong dịch tễ
  - » Phần mềm quản lý số liệu
  - » Phần mềm thống kê
- **Sử dụng** các phần mềm trong phân tích dịch tễ học cơ bản
  - » Phân tích đơn biến
  - » Phân tích đa biến
- Làm quen với phần mềm R

Chương trình

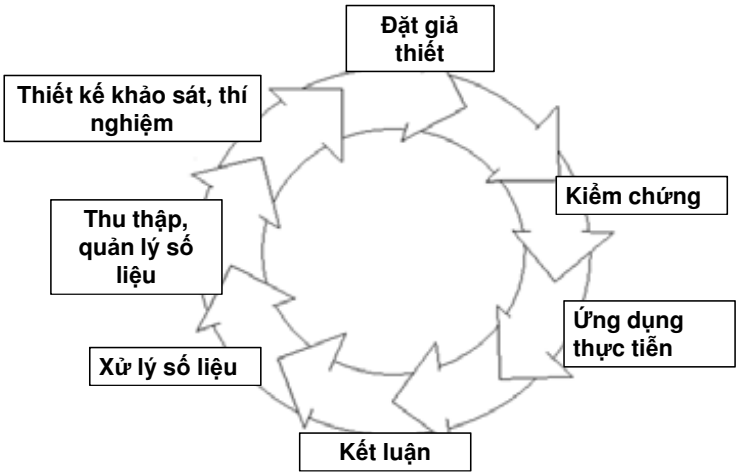
- **Ngày 1**
  - Giới thiệu
  - Quản lý số liệu bằng EpiData
  - Giới thiệu phần mềm thống kê STATA
- **Ngày 2**
  - Các nghiên cứu dịch tễ
  - Các đại lượng sử dụng trong nghiên cứu dịch tễ
  - Dùng STATA để tính các đại lượng dung trong nghiên cứu dịch tễ
  - Phân tích đa biến bằng Logistic
  - Bài tập
- **Ngày 3**
  - Thảo luận về bài tập
  - Giới thiệu phần mềm R
  - Thiết lập bản đồ nguy cơ bằng R
  - SaTScan để xác định cluster trên bản đồ dịch tễ

http://dichte.jimdo.com/

Ngày 1

Quản lý số liệu điều tra dịch tễ

Sơ đồ nghiên cứu



Mục tiêu

Sau khi hoàn thành phần thực tập này, các AC sẽ:

1. Hiểu được các thành phần của database
2. Biết cách thiết kế flat-file database bằng cách sử dụng EpiData
3. Hiểu được các nguyên tắc của databases
4. Biết được các phương pháp khác nhau để import data vào các phần mềm thống kê

Định nghĩa

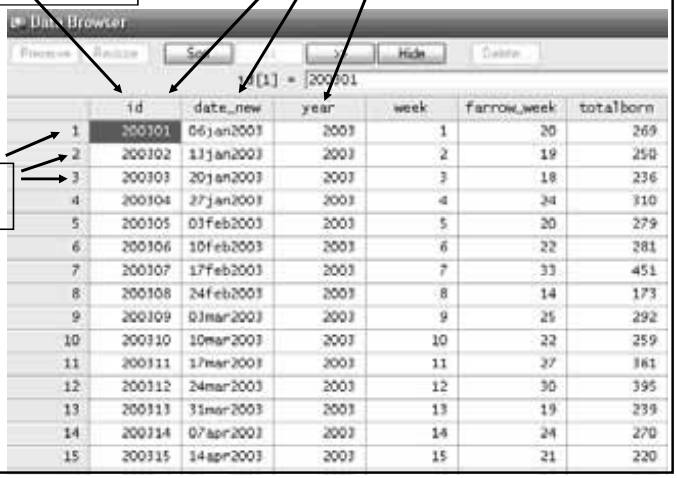
- Hệ thống quản lý Database
  - Database software
- Database
  - "Collection of data stored in some organized fashion" = "tập hợp các dữ liệu chứa trong những dạng có tổ chức"

Các hình thức lưu trữ data

Biến (Variables = Fields)

Mã xác định (Unique identifier = primary key, id)

Dữ liệu quan sát (Records)



	id	date_new	year	week	farrow_week	totalborn
1	200101	06jan2003	2003	1	20	269
2	200102	11jan2003	2003	2	19	250
3	200103	20jan2003	2003	3	18	236
4	200104	27jan2003	2003	4	24	310
5	200105	03feb2003	2003	5	20	279
6	200106	10feb2003	2003	6	22	281
7	200107	17feb2003	2003	7	33	451
8	200108	24feb2003	2003	8	14	173
9	200109	03mar2003	2003	9	25	292
10	200110	10mar2003	2003	10	22	259
11	200111	17mar2003	2003	11	27	361
12	200112	24mar2003	2003	12	30	395
13	200113	31mar2003	2003	13	19	239
14	200114	07apr2003	2003	14	24	270
15	200115	14apr2003	2003	15	21	220

Các bước trong quản lý dữ liệu

1. Định nghĩa biến
2. Tạo cấu trúc database và data dictionary
3. Kiểm tra quản lý dữ liệu trước khi bắt đầu thu thập số liệu
4. Nhập liệu và xác định những sai sót
5. Ghi chú những thay đổi
6. Định kỳ back up dataset
7. Tạo dataset để phân tích
8. Lưu trữ database ban đầu và database cuối cùng và file phân tích

Định nghĩa biến: Tên biến

- Tên gợi ý nghĩa của biến
- Số ký tự: 8 max
- Cố định nguyên tắc đặt tên
  - Ví dụ (date0, date1…; blood0, blood1…)
  - Khọng được bắt đầu bằng dấu #, @, …
- Dán nhãns:
  - Datasets
  - **Variables**
  - Variable values

Định nghĩa biến : Định dạng và giới hạn cho phép

- Data dictionary
- Đối với mỗi biến
  - Type
  - Format
    - Free text (medication A)
    - Binary variable (medication A yes or no)
  - Permissible values
- Nguyên tắc
  - E.g. 0 = negative
- Missing value

Variable	**Example**	**Variable/ storage type**
Continuous	Weight	Numerical (many)
Continuous	Time	Numerical
Interval	Body condition	Numerical/String*
Binary	Pos/Neg	Numerical
Nominal	Eye color	Numerical/String*
Free text	Open ended question. Names	String
*Although entered as numerical: will be analyzed as categorical		

Một số dạng database

Dataset format	**Database created by:**
All data in one spreadsheet	Researcher
All data in one table of a database created by data management system	Researcher
All data in multiple tables of a database (relational or not)	Researcher
All data in multiple tables of a relational database	Database professionals
Queried (Filtered data) from multiple tables of a relational database	Database professionals
Data from multiple tables and multiple sources (with or without referential integrity)	Different

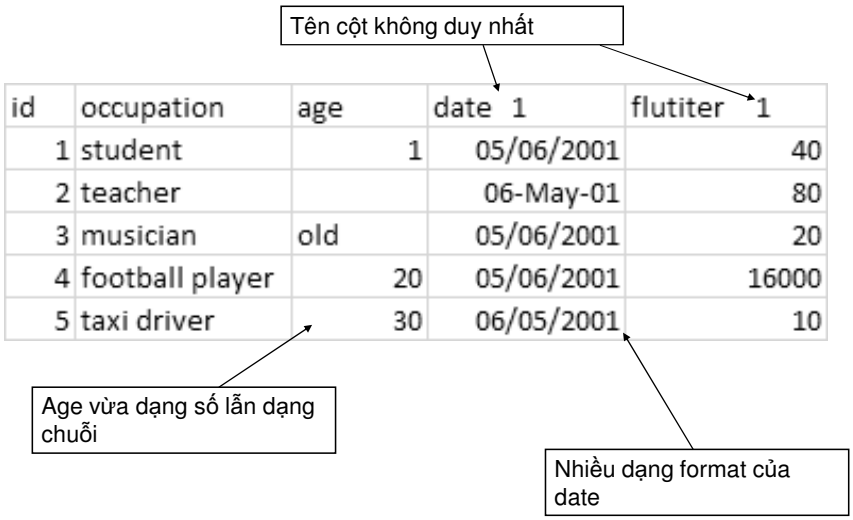
Databases

Health sciences terminology	Computer science terminology
Flat-file databases	Flat-file database
Relational databases	Hierarchical databases
	Network databases
	Relational databases
	Object-oriented databases

Flat-file database

- Tất cả thông tin trong 1 bản
- Data chứa trong các dạng
  - Spreadsheets
  - One table in a database
- EpiData Entry
- Giới hạn sử dụng trong database phân nhánh

Tránh sử dụng spreadsheets



id	occupation	age	date 1	flutiter 1
1	student	1	05/06/2001	40
2	teacher		06-May-01	80
3	musician	old	05/06/2001	20
4	football player	20	05/06/2001	16000
5	taxi driver	30	06/05/2001	10

Xuất/nhập data vào các statistical software

- Xuất data từ database để nhập vào các phần mềm thống kê
- EpiData to (Stata, SPSS, SAS), data and labels
- Any database – .txt or .csv file – stats package

Giới thiệu phần mềm

EpiData

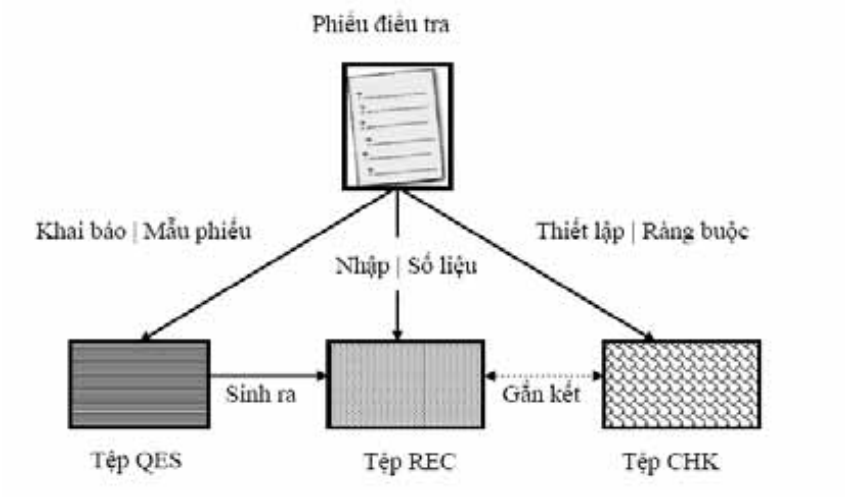
EPIDATA là gì

- EpiData là phần mềm hỗ trợ nhập và quản lý số liệu, được lập trình bởi Bác sĩ Jens M.Lauritsen, người Đan Mạch.
- Phần mềm này đã được sử dụng lần đầu tiên cho một nghiên cứu dịch tễ học "Phòng chống tai nạn".
- Ý tưởng của người phát triển phần mềm EpiData là việc tạo ra một phần mềm nhập liệu miễn phí, giao diện người dùng thân thiện, dễ sử dụng

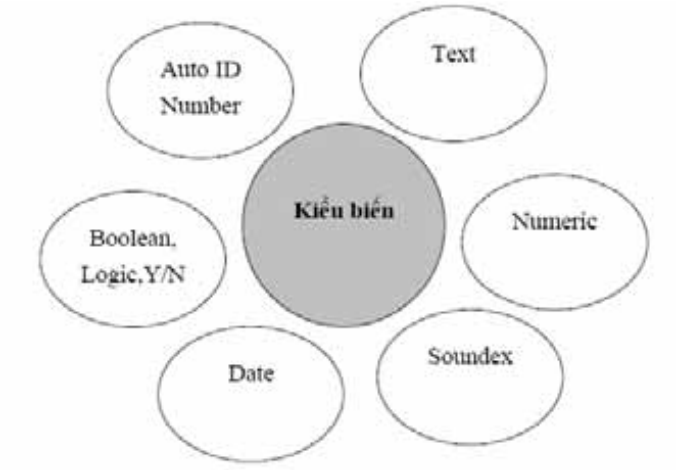
- EpiData là sản phần hoàn toàn miễn phí, người sử dụng có thể tải chương trình cài đặt từ trang Web http://www.epidata.dk.
- EpiData có thể chạy trên các máy tính cài đặt hệ điều hành Microsoft Windows hoặc Macintosh.
- Epidata có thể xuất số liệu sang nhiều dạng khác nhau để sử dụng cho phân tích số liệu bằng các phần mềm như Stata, Spss, .v.v.

Dạng file trong epidata

- File dạng *.qes
  - File thiết kế bản questionaire
- File dạng *.rec
  - File nhập và lưu trữ dữ liệu
- File dạng CHK


Phiếu điều tra
Khai báo | Mẫu phiếu
Nhập | Số liệu
Thiết lập | Ràng buộc
Sinh ra
Gắn kết
Tệp QES
Tệp REC
Tệp CHK

Kiểu dữ liệu


Auto ID
Number
Text
Kiểu biến
Numeric
Boolean,
Logic,Y/N
Date
Soundex

Kiểu ID number

- Chuỗi định dạng là <IDNUM>
- Một trường số liệu được khai báo kiểu ID number thì giá trị số liệu của trường sẽ được tự động nhập khi nhập số liệu. Người sử dụng không được nhập giá trị cho trường này.
- Vd: IDX So thu tu phong van <IDNUM>

Kiểu Numeric

- Chuỗi định dạng sử dụng kí tự #, ví dụ ###, hoặc ###.###, hoặc ########, hoặc ##.#### .v.v.
- Trường được khai báo kiểu số chỉ chấp nhận số liệu nhập vào ở dạng số.
- Độ rộng của trường được xác định bằng số kí tự # được khai báo.
- Kích cỡ lớn nhất số liệu nhập vào một trường có kiểu số là 14 chữ số gồm cả ký tự (“.”)ngăn cách phần số nguyên và phần thập phân với số thập phân.

Kiểu Text

- Chuỗi định dạng là sử dụng ký tự "_" hoặc <E >
- Chuỗi văn bản nhập vào có thể gồm các ký tự a, b, c, … và kể cả các chữ số.
- Độ rộng của trường lớn nhất là 80 kí tự.
- Khi khai báo mỗi dấu "_" tương ứng với khai báo cho một kí tự.
- Vd: V2 Ho ten _____

- *Kiểu Upper-case text*
  - Chuỗi định dạng là <A>, hoặc <A >
  - Trường được khai báo kiểu Upper-case text thì số liệu nhập vào trường này được hiểu là dạng văn bản và được tự động chuyển sang dạng kí tự viết hoa.
  - Độ rộng của trường tương ứng với số kí tự "trống" (dấu cách) giữa hai dấu "<" và ">".

Kiểu Boolean

- Đây là kiểu dữ liệu logic.
- Trường được khai báo kiểu này chỉ chấp nhận giá trị Y hoặc N (cũng có thể chấp nhận số 0 hoặc 1)
- Chuỗi định dạng là <Y>

Kiểu Date

- Chuỗi định dạng là <dd/mm/yyyy>, hoặc <mm/dd/yyyy>, hoặc <yyyy/mm/dd>

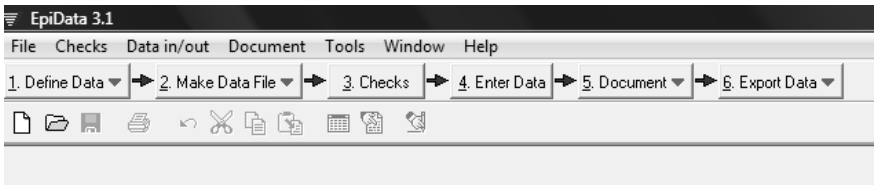
- *Kiểu today's date*
  - Chuỗi định dạng là <today-dmy>, hoặc <today-mdy>, hoặc <today-ymd>
  - Một trường được khai báo kiểu Today's date sẽ được tự động điền vào giá trị ngày hiện tại (ngày của máy tính) khi nhập liệu.

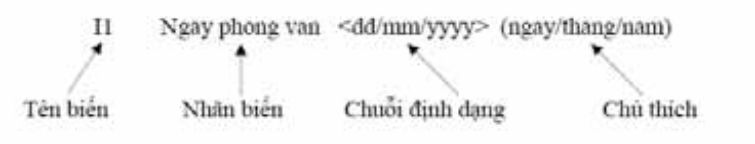
Kiểu soundex

- Kiểu Soundex là kiểu dữ liệu mã hóa. Số liệu nhập vào trường này sẽ được Epidata tự động mã hóa (chuyển sang một giá trị khác) theo quy luật mã hóa của Epidata trước khi lưu vào cơ sở dữ liệu
- Chuỗi định dạng là <S >

Cài đặt Epidata

- Mở cửa sổ trình duyệt Internet Explorer vào trang web www.epidata.dk, vào mục Download (get files).
- Chọn liên kết Epidata Entry để nhảy đến phần Epidata Entry
- Chọn mục Complete setup để bắt đầu tải tệp chương trình cài đặt về máy tính.
- Tiến hành install





Giới thiệu phần mềm

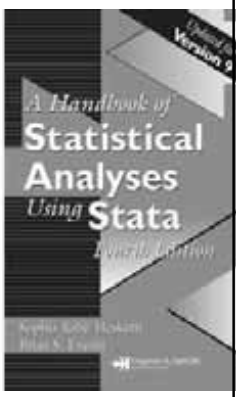
STATA

- Phần mềm xử lý thống kê của StataCorp
- Ý nghĩa là "statistics" and "data“
- Được sử dụng ngày càng rộng rãi vì khả năng phân tích mạnh và dễ sử dụng, bên cạnh đó có thể lập trình và giá thành rẻ
- Nhiều chức năng phân tích ứng dụng trong dịch tễ

Tài liệu Tham khảo

- http://www.ats.ucla.edu/stat/stata/
- **Handbook of Statistical Analyses Using Stata**

  By Sophia Rabe-Hesketh, Brian S. Everitt



Cài đặt STATA

Các thành phần của STATA

- Command
- Result
- Review
- Variables
- Data editor

Một số vấn đề

- Thao tác của Stata có thể được thực hiện thông qua 2 kiểu
  - graphical user interface (GUI)
  - command line
- Do-file
- log

Nhập liệu

1. Nhập trực tiếp
2. Nhập từ gián tiếp

Nhập trực tiếp

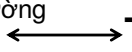
- Tạo biến bằng lệnh genenate = gen
- Nhập số liệu bằng lệnh edit
- Label biến

- Tương tự thực hiện thao tác bằng GUI

Nhập gián tiếp

Các phầm mềm quản lý số liệu

Phầm mềm thông thường
- Excel, Access



Text file

Phầm mềm chuyên dùng
- Epidata, survey toolbox

Thao tác

- Mở file lab0.txt bằng Excel, lưu lại thành lab1.xls và lab1.csv
- Import vào Stata
- Thêm biến "barn" với giá trị "1" cho id từ 1-50 (ngoại từ id=2) và còn lại là 2
  - **generate barn = 1**
  - **replace barn =2 if id > 50 | id ==2**
    - Ghi chú: "&" nghĩa là "And", "|" nghĩa là "Or", v à "~=" hay "!=" à "not equal"; "==" được sử dụng trong lệnh if
- **Dùng lệnh save để lưu thành file lab2.dta và lệnh clear để đóng.**
- **Xem nội dung file lab_merge.dta**
- **Nối lab2 và lab_merge.dta**
  - merge id using "C:\... \merge.dta", unique
  - GUI: Data >combine datasets > merge two datasets >

Một số lệnh khác

- **Codebook**
- **summarize và bysort**
  - Summarize
  - summarize, detail
  - summarize weight, detail
  - bysort sex: summarize weight

  - ***GUI:*** *Data > Describe data > Summary statistics > Nhập weight cho "variables" và ấn"Repeat commands by group" chọn"sex"*

- tabstat weight, by(sex) columns(variables)
- ***GUI:*** *Statistics > Summaries, tables, and tests > Tables > Tables of summary statistics (tabstat) > chọn weight cho "variable" và "group statistics" bằng sex*.

Biểu đồ

- histogram weight, percent by(barn)
- ***GUI:*** *Graphics > Histogram > chọn "weight" là biến chính(Main tab), "percent" cho trục y (Main tab), và "barn" như là biến phân lọai (By tab)*

Vấn đề khác

- Lưu graph
- log , copy cửa sổ result
- cd "C:\...\“ để cài đặt thư mục hiện hành
- Thực tập do-file
  - Dùng “*” để phân biệt câu không lệnh
  - Dùng “/” để báo hiệu câu lệnh xuống dòng

Ngày 2

Thực hành

DỊCH TỄ HỌC
Cơ Bản

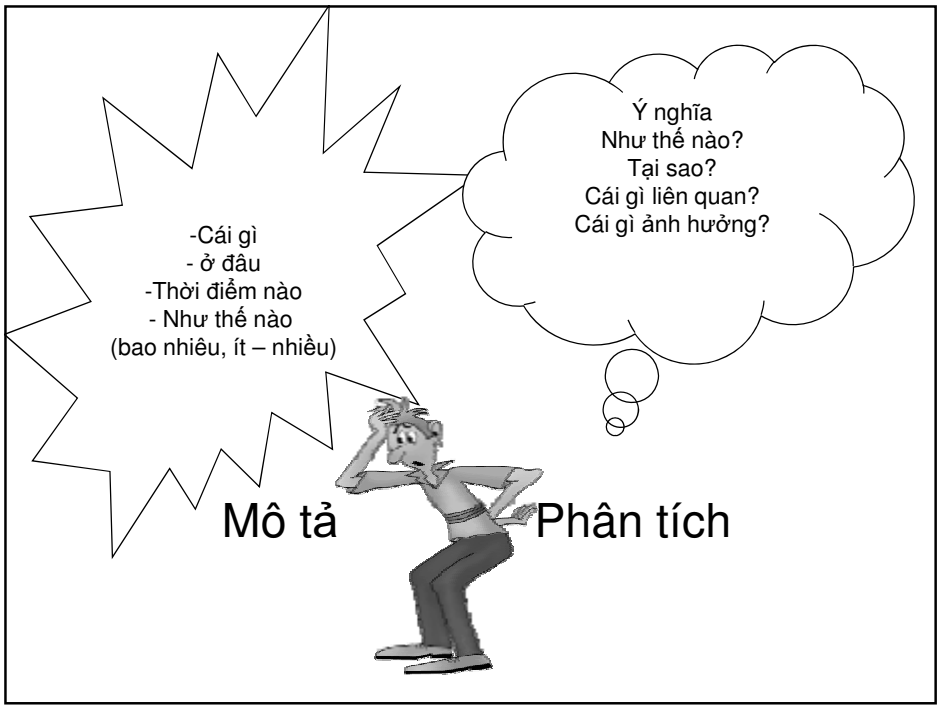
- Dịch tễ học là môn học nghiên cứu về mối liên quan giữa tác nhân gây bệnh, yếu tố truyền lây, môi trường và vật chủ

- Theo Last (1995),
  Dịch tễ học là môn học ứng dụng thống kê và nhiều ngành khoa học khác để nghiên cứu về sự phân bố bệnh, các yếu tố liên quan đến bệnh trong một quần thể xác định. Ứng dụng trong việc xác định nguyên nhân gây bệnh và kiểm soát dịch bệnh

Các nghiên cứu dịch tễ học

- **Dịch tễ học mô tả** (descriptive epi. Study)
  - Ai? Cái gì? ở đâu? Khi nào?

- **Dịch tễ học phân tích** (analytic epi. Study)
  - Như thế nào? Tại sao
    + thí nghiệm
    + nghiên cứu quan sát



MỘT SỐ KHÁI NIỆM TRONG DỊCH TỄ HỌC MÔ TẢ

- **1. Tỉ lệ bệnh** (prevalence)
- **2. Tỉ lệ mắc bệnh** (incidence)

  Tỷ lệ mắc bệnh tích lũy (Cumulative Incidence: CI)

  Tốc độ mắc bệnh (Incidence Density Rate: IR)

1. Tỷ lệ bệnh (prevalence)

- *Tỷ lệ bệnh, tỷ lệ nhiễm.*
- *Là số con thú có cùng tính chất đang khảo sát (bệnh, nhiễm bệnh, mang trùng, có rối loạn bất thường về sức khỏe...) trong một quần thể tại một thời điểm nhất định chia cho tổng số thú trong quần thể đó.*
- *Đại lượng này thường được tính theo phần trăm.*

$$P\ (\%) = \frac{\text{Số thú mắc bệnh x 100}}{\text{Tổng số thú trong quần thể tại một thời điểm nhất định}}$$

$$P = p \pm 1{,}96 \times \sqrt{p(1-p)/n}$$

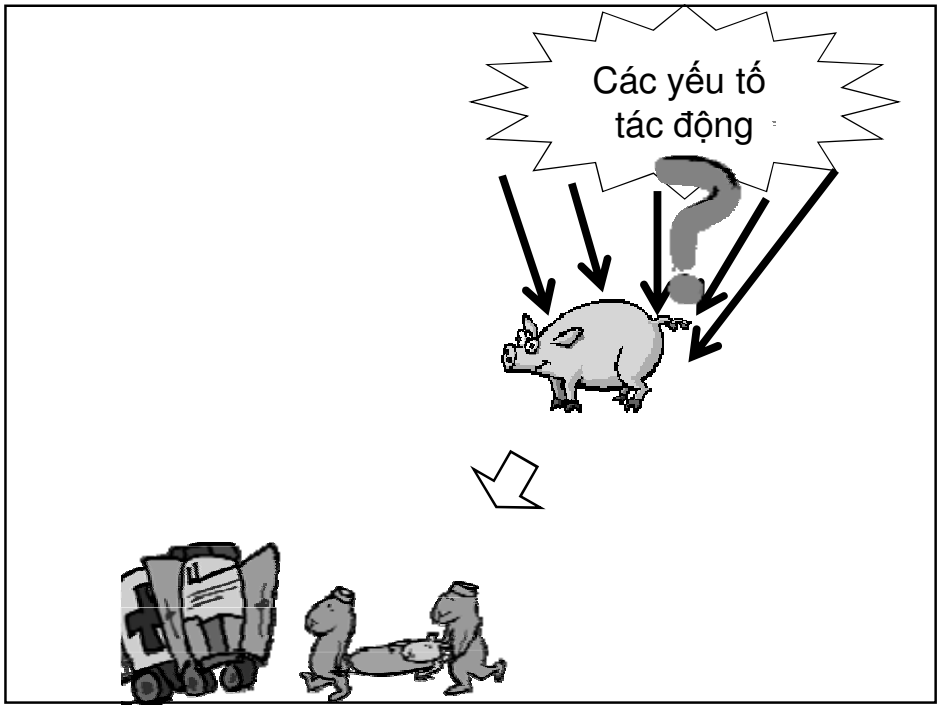
- ***Tỷ lệ mắc bệnh tích lũy*** (CI) là tỷ lệ giữa số thú mắc bệnh trong một khoảng thời gian nhất định và số con thú khỏe có nguy cơ mắc bệnh trong quần thể ở đầu thời gian khảo sát.
- Như vậy CI là một đại lượng đặc trưng cho nguy cơ mắc bệnh của quần thể trong thời gian khảo sát.
- Đây là đại lượng thường được dùng trong các nghiên cứu dịch tễ học phân tích.
- CI có giá trị từ 0 đến 1.

Thực hành với STATA

- Xác định tỉ lệ bệnh và tỉ lệ mắc bệnh, so sánh 2 quần thể
  - prtesti N1 p1 N2 p2
  - GUI: Statistics>summaries, tables, test> classical tests > Two groups proportion test

DỊCH TỄ HỌC PHÂN TÍCH

(Mối quan hệ giữa yếu tố nguy cơ và bệnh)



- ***Nguy cơ*** *là khả năng có thể mắc một bệnh nào đó, nguy cơ được định nghĩa là xác suất xuất hiện một biến cố có liên quan đến sức khỏe của mỗi cá thể hay quần thể.*

- *Trong khi đó bất kỳ yếu tố nào, thuộc bản chất nào (lý học, hoá học, sinh học, di truyền, xã hội...) góp phần vào việc làm cho cơ thể đang khoẻ mạnh trở nên mắc bệnh thì yếu tố đó được gọi là* ***yếu tố nguy cơ****.*

Đo lường mối quan hệ để đánh giá yếu tố nguy cơ

- **Bệnh** vừa có nghĩa là bệnh nhưng đồng thời bao gồm luôn các vấn đề sức khỏe được quan tâm.
- **Nhóm phơi nhiễm** (tiếp xúc yếu tố nguy cơ) (E+: exposed group) là nhóm được khảo sát sự hiện diện của bệnh, trong đó các cá thể đều có chung yếu tố nguy cơ. Ví dụ nhóm người hút thuốc lá.
- **Nhóm không phơi nhiễm** (không tiếp xúc yếu tố nguy cơ) (E-: non-exposed group) là nhóm không có tính chất, hoặc không mang yếu tố nguy cơ. Ví dụ nhóm người không hút thuốc.

Các giá trị tính

- tỷ số nguy cơ hay nguy cơ tương đối (relative risk hay risk ratio) **(RR)**
- tỷ số của tốc độ độ bệnh (IRR: incidence rate ratio)
- tỷ số chênh (odd ratio) **(OR)**

Các nghiên cứu dịch tễ phân tích

- Nghiên cứu đoàn hệ (cohort study)
- Nghiên cứu bệnh chứng (case-control study)
- Nghiên cứu cắt ngang

Cách xác định tỷ số nguy cơ hay nguy cơ tương đối (relative risk hay risk ratio) (RR)

		Yếu tố khảo sát		
		Phơi nhiễm (E+)	*Không phơi nhiễm (E-)*	Tổng
Kết quả	*Bệnh*	a	b	a + b
	Không bệnh	c	d	c + d
	Tổng	a + c	b + d	N

$$RR = \frac{P(D+/E+)}{P(D+/E-)} = \frac{a/(a+c)}{b/(b+d)}$$

- RR<1 yếu tố nguy cơ có mối quan hệ nghịch, tức là mối quan hệ bảo vệ chống lại bệnh (ví dụ như vắc-xin)
- RR=1 yếu tố nguy cơ không có liên quan đến bệnh
- RR>1 yếu tố nguy cơ có liên quan đến bệnh

Tỷ số chênh (odd ratio)

•"Chênh" (odd) được định nghĩa như tỷ phần giữa 2 đặc điểm trong một nhóm. Ví dụ, trong một nhóm thú gồm n con trong đó có x con bệnh, chỉ số odd của bệnh trong nhóm là x/(n-x).
•Tỷ số chênh (OR) là tỷ số giữa chỉ số odd của nhóm thú phơi nhiễm và chỉ số odd của nhóm không phơi nhiễm.

		Yếu tố khảo sát		
		Phơi nhiễm (E+)	*Không phơi nhiễm (E-)*	Tổng
Kết quả	*Bệnh*	a	b	a + b
	Không bệnh	c	d	c + d
	Tổng	a + c	b + d	N

$$OR = \frac{odd(D+/E+)}{odd(D+/E-)} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

* Đánh giá
OR = 1 → không có mối liên quan
OR > 1 → có mối liên quan; yếu tố nguy cơ có thể làm tăng tỉ lệ bệnh
OR < 1 → mối liên quan dạng bảo vệ

Thực hành tính RR, OR bằng Stata

- Mở file lab_day2.dta
- Dùng lệnh cc disease expose cho case-control study và cs cho cohort study để tính các thông số dịch tễ
- GUI: **Statistics > Epi. And related > Tables for epidemiologist >**

•<u>Yếu tố gây nhiễu (confounder)</u>

Trong các nghiên cứu dịch tễ học mô tả lẫn phân tích, người ta thường quan tâm đến yếu tố nhiễu mà làm cho kết quả khảo sát giữa các quân thể có thể bị ảnh hưởng. ví dụ như sự phân bố về tuổi giữa 2 quân thể khác nhau làm cho tỉ lệ bệnh của chúng khác nhau nhưng về bản chất là tỉ lệ bệnh giữa hai quần thể này là tương đương.

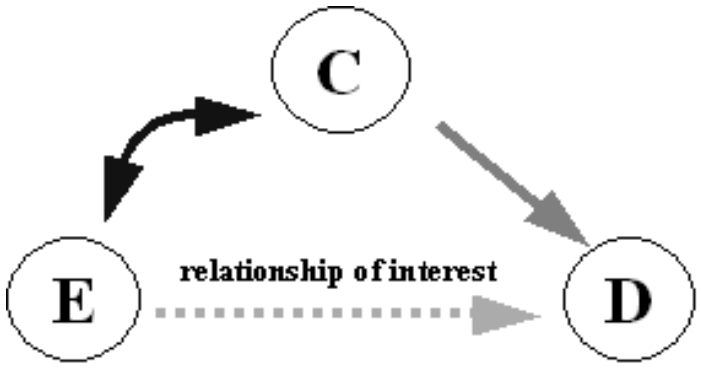


Figure 1

Để khắc phục tình trạng này, người ta thường dùng các biện pháp sau:

•lấy mẫu dạng bắt cặp (dùng trong dịch tễ phân tích) (matching)

•phân tầng (stratification)

•thêm biến trong mô hình (dùng trong thiết lập tương quan tuyến tính)

•hiệu chỉnh (dùng trong dịch tễ mô tả)

•Đánh giá yếu tố gây nhiễu:

Để xác định xem yếu tố X nào đó có phải là yếu tố nhiễu trong mối quan hệ giữa E và D, có nhiều cách đánh giá.

- Thông thường có thể đánh giá bằng cách tính phần trăm giữa chỉ số liên quan RR hay OR chưa hiệu chỉnh và đã hiệu chỉnh. Mức khác biệt trên 10% cho thấy khả năng X là yếu tố nhiễu.

- Bên cạnh đó, trong thống kê có thể dùng Breslow-Day tests để đánh giá mức độ tương đồng của hai giá trị liên quan hiệu chỉnh và chưa hiệu chỉnh.

Thực hành đánh giá yếu tố nhiễu bằng STATA

- cs disease barn, by(sex)
- GUI: **Statistics > Epi. And related > Tables for epidemiologist >**

sex	RR	[95% Conf.	Interval]	M-H Weight
female	2.708333	1.257563	5.832766	2.88
male	2.30303	1.215707	4.362851	2.64
Crude	2.405229	1.49478	3.870219	
M-H combined	2.514493	1.512955	4.179024	

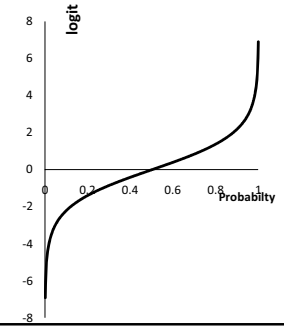
Test of homogeneity (M-H) chi2(1) = 0.109 Pr>chi2 = 0.7417

PHÂN TÍCH ĐA BIẾN BẰNG LOGISTIC

$$g(x) = \ln\left(\frac{p}{1-p}\right) = B_o + B_1 X$$

$$p = \frac{e^{B_o + B_1 X}}{1 + e^{B_o + B_1 X}} \qquad odds = \frac{p}{1-p} = e^{B_o + B_1 X}$$

$$or = \frac{e^{B_o + B_1 X}}{e^{B_o}} = e^{B_1}$$

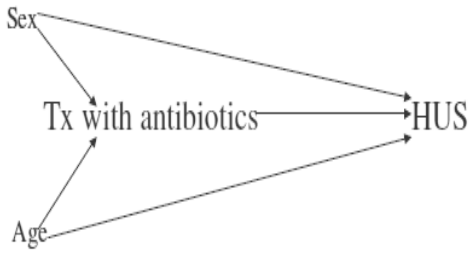


Ln{p/(1-p)} = βo + ΣβjXj

Trong đó p là xác suất để xảy ra tính chất cần xác định (chẳng hạn như xác suất có bệnh) và X là các yếu tố quan sát có ảnh hưởng. Tỉ số p/(1-p) được gọi là Odd

=> Odd= $e^{\beta o + \Sigma \beta j Xj}$

Ví dụ



Sex: Female (1) / Male (0) Age: Continuous Tx: Yes (1) / No (0) HUS: Yes (1) / No (0)

HUS: Hemolytic-uremic syndrome

```
logit HUS Tx age sex

     Iteration 0:   log likelihood = -449.93213   Log (L) of null model
     Iteration 1:   log likelihood = -407.46919
     Iteration 2:   log likelihood = -403.75126
     Iteration 3:   log likelihood = -403.67715
     Iteration 4:   log likelihood = -403.67709

     Logit estimates                                   Number of obs   =        869
                                                       LR chi2(3)      =      92.51
                                                       Prob > chi2     =     0.0000
     Log likelihood = -403.67709                       Pseudo R2       =     0.1028

                                          = Coef. / Std. Err.
     ------------------------------------------------------------------------------
             HUS |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
     ------------+-----------------------------------------------------------------
              Tx |   1.140377   .2204085     5.17   0.000     .7083841    1.572369
             age |  -.0346741   .0051351    -6.75   0.000    -.0447387   -.0246096
             sex |   .1666718   .1782025     0.94   0.350    -.1825987    .5159423
           _cons |  -.8829411   .1572963    -5.61   0.000    -1.191236    -.574646
                                          Wald test          CI didn't cover "0"
```

```
logit HUS Tx age sex, or

Iteration 0:   log likelihood = -449.93213   Log (L) of null model
Iteration      ..... Everything is tthe same
------------------------------------------------------------------------------
         HUS | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Tx |   3.127947   .6894259     5.17   0.000     2.030707     4.81805
         age |   .9659201   .0049601    -6.75   0.000     .9562474    .9756908
         sex |   1.181366   .2105225     0.94   0.350     .8331024    1.675216
------------------------------------------------------------------------------
```

NO _cons — exp(coef.) — CI did cover "1"

- Đánh giá
  - Điều trị bằng kháng sinh làm tăng odd của HUS 3.13 lần ($e^{1.14}$).
  - Ex: odds của HUS bằng 0.97 lần cho mỗi tuổi lớn, hoặc 0.73 lần cho khoảng 9 tuổi chênh lệch.

$$\log odds(x_1, x_2) = (x_2 - x_1) * B_1 = (10-1) * (-0.0347) = -0.312$$

$$e^{-0.312} = 0.732$$

$$OR(x_1, x_2) = OR^{(x_2 - x_1)} = 0.966^{(10-1)} = 0.732$$

Độ ý nghĩa của mô hình

1. Wald tests
   - Tỉ số giữa coefficient (log odds scale) và standard error.
   - Có phân phối chuẩn z
   - Tests coefficient (log odds scale) bằng "0" (Ho)
   - Không nên quá phụ thuộc vào test này

2. Likelihood ratio test

- G^2_o so sánh mô hình full và mô hình null. Có phân bố chi-sq. Tương tự F test trong ANOVA
- G^2 = 2 (ln *L* full – ln *L* red). So sánh mô hình full và mô hình reduced (nested). = Partial F test

$$G_o^2 = 2\ln\frac{L}{L_o} = 2(\ln L - \ln L_o) = 2(-403.677 - (-449.932)) = 92.51$$

$$\text{df} = p - 1 = k = 3 \quad \Longrightarrow \quad p = 0.0000$$

$$G^2 = 2(\ln L_{\text{full}} - \ln L_{\text{red}}) = 2(-403.677 - (-404.116)) = 0.878$$

$$\text{df} = 1 \qquad p = 0.349$$

```
logit HUS  Tx age sex        Full model

est store a                  Name it "a"

logit HUS  Tx age            Reduced model

Iteration 0:   log likelihood = -449.93213
Iteration 1:   log likelihood = -407.87164
Iteration 2:   log likelihood = -404.18915
Iteration 3:   log likelihood = -404.11649
Iteration 4:   log likelihood = -404.11642

Logit estimates                                   Number of obs   =        869
                                                  LR chi2(2)      =      91.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -404.11642                       Pseudo R2       =     0.1018

------------------------------------------------------------------------------
         HUS |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Tx |   1.146267   .2201535     5.21   0.000     .7147738    1.577759
         age |  -.0341116   .0050945    -6.70   0.000    -.0440967   -.0241265
       _cons |  -.8005654   .1291564    -6.20   0.000    -1.053707   -.5474234
------------------------------------------------------------------------------

est store b                  Name it "b"
lrtest a b                   compare "a" and "b"

likelihood-ratio test                             LR chi2(1)  =      0.88
(Assumption: . nested in LRTEST_0)                Prob > chi2 =    0.3486

                                                  Non-sig means reduced and
                                                  full model are the same

                                                  => better use reduced model
```

Chọn lựa mô hình tối ưu

- estat ic để xem AIC của mô hình; AIC càng nhỏ càng tốt
- Mô hình càng đơn giản càng tốt

lincom

Female (5 years):	B1 (Tx)	B2 (Age)	B3 (Sex)	B4 (Tx*Sex)
Tx=1	1	5	1	1
Tx=0	0	5	1	0
Difference:	1	0	0	1

```
lincom 1*Tx + 1*sex_tx_int
 ( 1)  Tx + sex_tx_int = 0
------------------------------------------------------------------------------
         HUS |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.252046   .2892456     4.33   0.000     .6851347    1.818957
------------------------------------------------------------------------------
```

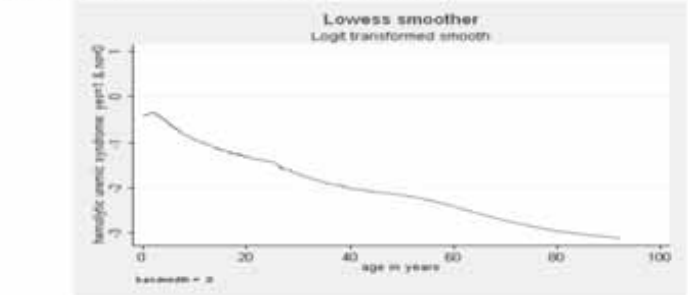
Comparison:		B1 (Tx)	B2 (Age)	B3 (Sex)	B4 (Tx*Sex)
Tx=1	Age=5	1	5	1	1
Tx=0	Age=10	0	10	1	0
Difference:		1	-5	0	1

```
lincom 1*Tx + 1*sex_tx_int + (-5*age),or
 ( 1)  Tx - 5 age + sex_tx_int = 0
------------------------------------------------------------------------------
         HUS | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   4.159747   1.207017     4.91   0.000     2.355468    7.346095
------------------------------------------------------------------------------
```

Sử dụng biến liên tục trong mô hình

- Đánh giá xu hướng bằng lintrend hay lowess hay lintrend Y X, groups (#) plot(log)

lowess HUS age, logit



Chuyển biến liên tục thành biến phân loại

- Không cần giả định xu hướng trong mô hình
- Dễ thực hiện
- Tạo các biến giả (dummy variables)
- Đánh giá theo từng loại

- Thao tác: có thể làm thủ công hay tự động

```
logit HUS  Tx  age_2 age_3 age_4 sex
  Iteration 0:   log likelihood = -449.93213
  Iteration 1:   ---------------------------
  Logit estimates                                   Number of obs   =        869
                                                    LR chi2(5)      =     104.24
                                                    Prob > chi2     =     0.0000
  Log likelihood = -397.81303                       Pseudo R2       =     0.1158

  ------------------------------------------------------------------------------
           HUS |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
            Tx |   1.093117   .2228701     4.90   0.000     .6562996    1.529934
         age_2 |  -.8014851   .2205305    -3.63   0.000    -1.233717   -.3692533
         age_3 |  -.5365804    .263461    -2.04   0.042    -1.052954   -.0202064
         age_4 |  -.7384187   .3298555    -2.24   0.025    -1.384924   -.0919138
           sex |    .213418   .1814515     1.18   0.240    -.1422204    .5690565
         _cons |  -.7539106   .1684886    -4.47   0.000    -1.084142   -.4236791
```

  - Thay đổi giá trị baseline
    - char age_quart [omit] 4
    - xi: i. age_quart
- Predict p

Bài tập

Ngày 3

Phân tích số liệu
và biểu đồ bằng

- Do Ross Ihaka và Robert Gentleman - Trường đại học Auckland, New Zealand phát hoạ 1996
- R là một phần mềm sử dụng cho phân tích thống kê và vẽ biểu đồ
- R là ngôn ngữ máy tính đa năng, có thể sử dụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí (recreational mathematics), tính toán ma trận (matrix), đến các phân tích thống kê phức tạp

Tải R xuống và cài đặt vào máy tính

- Tải chương trình: Tài liệu cần tải về, tùy theo phiên bản R và số phiên bản (R-2.9.1-win32.exe)

- Tải package
- http://cran.R-project.org
- Địa chỉ để tải các package vẫn là: http://cran.r-project.org, rồi bấm vào phần “Packages” xuất hiện bên trái của mục lục trang web
- Các package này có thể cài đặt trực tuyến bằng cách chọn Install packages trong phần packages của R
- Lệnh library() để biết các package đã download

"Văn phạm" R

đối tượng <- hàm(thông số 1, thông số 2, …, thông số *n*)

- Chẳng hạn như:

```
reg <- lm(y ~ x)
```

Một số kí hiệu hay dùng trong R

x == 5	x bằng 5
x != 5	x không bằng 5
y < x	y nhỏ hơn x
x > y	x lớn hơn y
z <= 7	z nhỏ hơn hoặc bằng 7
p >= 1	p lớn hơn hoặc bằng 1
A & B	A và B (AND)
A | B	A hoặc B (OR)
!	Không là (NOT)

Với R, tất cả các câu chữ hay lệnh sau kí hiệu # đều không có hiệu ứng

Cách đặt tên đối tượng trong R

- Không giới hạn
- Phân biệt chữ hoa chữ thường
- Không nên có các dấu _ hay –
- Có thể dùng dấu "."
- Tránh dùng trùng với tên biến trong dữ liệu, tên hàm hay tên package
- Dấu $ phân cách giữa tên đối tượngvà tên biến trong đối tượng đó
  - Data
  - Data$age

Các dạng đối tượng

- Vector (vector)
- Ma trận (matrix)
- Bảng khung (dataframe)
- Mảng (array)

```
> weight<-c(1,2,3,4,5)
> c
> is.vector(weight)
> height<-c(1,3,4,6,8)
> plot(weight,height)

> Mx <-matrix(c(7,6,6,10,1,29,6,34,2,56,45,3,43,
23,1,3),nrow=4, byrow=T)
> colnames(Mx)<- c("pos","neg","up","down")
>  rownames(Mx)<- c("M","N","B","Z")
> Mx
```

Nhập bảng số liệu từ file text

```
MyData<-read.table(file.choose(), head=T)
head(MyData)

> save.image("C:\\Users\\....\Documents\\mydata")
> save(MyData,Mx, file=""C:\\Users\\....\Documents\\mydata")

ls()   # liệt kê các đối tượng trong file

> MyData<-read.csv(file.choose(), head=T)  # chọn file data_ex trong day3
> glm1 <- glm(MyData$mastitis ~ MyData$milking + MyData$dipping +
   MyData$drying ,family=binomial)
Summary(glm1)
```

Thực hành tạo bản đồ nguy cơ trong R

- **Nhập liệu file Akdat: tình hình bệnh giả dại trên heo của 482 hộ chăn nuôi tại 1 tỉnh**

```
> ak <- read.table(file.choose(),header=T)
> head(ak)
> names(ak) <- c("x", "y", "status")
> head(ak)
```

- Tạo dataset của bệnh

```
> akpos <- ak[ak$status == 1,]
> akposmap <- akpos[c(1,2)]
> head(akposmap)
```

- Tạo dataset của không bệnh

```
> akneg <- ak[ak$status == 0,]
> aknegmap <- akneg[c(1,2)]
> head(aknegmap)
```

- Tạo đường biên giới

```
x <- c(3.909, 5.662, 4.545, 6.934, 10.114, 15.202, 18.864,
   19.190, 21.734, 21.889, 23.797, 26.031, 31.119, 31.600,
   37.169, 38.131, 33.989, 34.935, 31.910, 24.759, 19.655,
   17.747, 14.892, 11.696, 10.750, 13.294, 12.503)

y <- c(22.832, 15.184, 8.978, 6.590, 5.643, 7.396, 10.096,
   12.795, 13.431, 16.146, 18.364, 18.364, 15.664, 21.870,
   20.598, 21.715, 26.322, 29.673, 31.907, 27.610, 31.582,
   34.762, 32.854, 33.490, 30.945, 28.712, 25.376)

ak.poly <- as.data.frame(cbind(x,y))
```

Tạo bản đồ phân bố

- Download package splancs sau đó load splancs

```
windows()
   par(pty="s", cex=1.2)
   pointmap(as.points(ak.poly), col=-1)
   points(as.points(akneg$x,akneg$y), pch=1, cex=0.7,
      col=1)
   points(akpos$x,akpos$y, pch=16, , cex=0.7, col=4)
   polygon(ak.poly, density=0)
```

Phương pháp “Kernel smothing” để tạo bản đồ nguy cơ

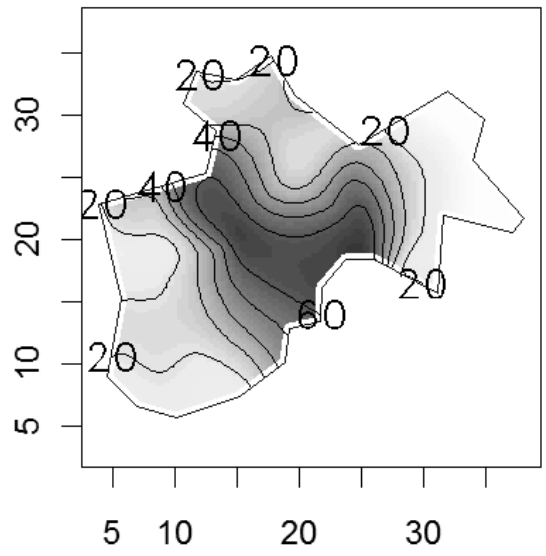
- Tạo file điểm riêng cho case và control

```
pts1 <- as.points(akposmap) # case population
pts2 <- as.points(aknegmap) # control population
pts3 <- as.points(rbind(akposmap,aknegmap)) # population at risk
```

Tạo bản đồ

```
riskmap <- kernrat(pts1, pts3, as.matrix(ak.poly),
   h1=4, h2=4, nx=118, ny=104)

windows()
par(pty="s", cex=1.5)
pointmap(as.points(ak.poly), col=-1)
image(riskmap$x, riskmap$y, -riskmap$z ,col =
   heat.colors(30), add=T)
```

```
polymap(ak.poly, col="white", dens=0, lwd=8,
   add=T)
polymap(ak.poly, col=4, dens=0, lwd=1, add=T)
contour(riskmap$x, riskmap$y, 100*riskmap$z,
add=T,levels=c(20,30,40,50,60),
   labels=c("20","","40","","60"),
lwd=1,
labcex=2,method="simple", vfont = c("sans serif",
   "bold"))
```



Xác định vùng trung tâm dịch bằng

SaTScan™

Software for the spatial, temporal, and space-time scan statistics

http://www.satscan.org/

Chuẩn bị data trên R

- Mở file AKdat

```
prv.data <- read.table(file.choose(),header=T)
head(prv.data)
prv.geo<-prv.data[,1:2]
prv.cas<-prv[prv.data$POS==1,]
prv.ctl<-prv.data[prv.data$POS==0,]
```

- Thêm thời gian

```
prv.cas$year<-rep(1997,length(prv.cas$POS))
prv.ctl$year<-rep(1997,length(prv.ctl$POS))
```

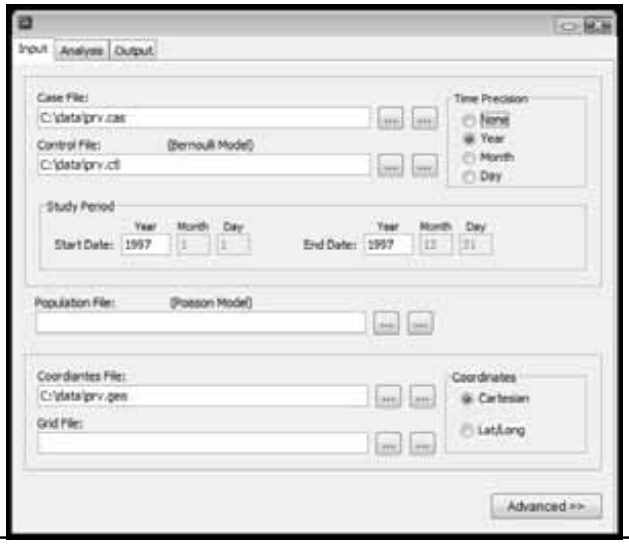
Lưu cđối tượng thành file text

```
write.table(prv.geo,"C:/data/prv.geo",row.names=TRUE,col.names=
    FALSE,quote=F)
head(prv.geo)

write.table(prv.cas[,3:4],"C:/data/prv.cas",row.names=TRUE,col.na
    mes=FALSE,quote=F)
head(prv.cas[,3:4])

prv.ctl$POS<-rep(1,length(prv.ctl$POS)) # đổi pos “0” thành “1”
write.table(prv.ctl[,3:4],"C:/data/prv.ctl",row.names=TRUE,col.nam
    es=FALSE,quote=F)
head(prv.ctl[,3:4])
```

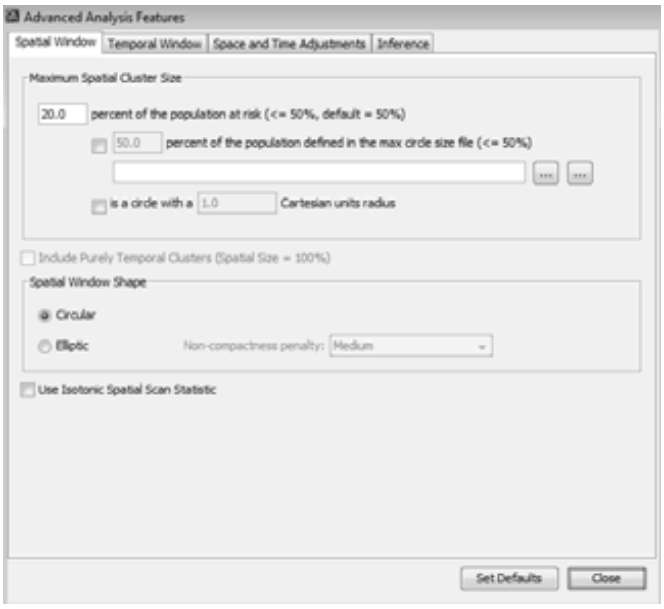
Khởi động SaTScan

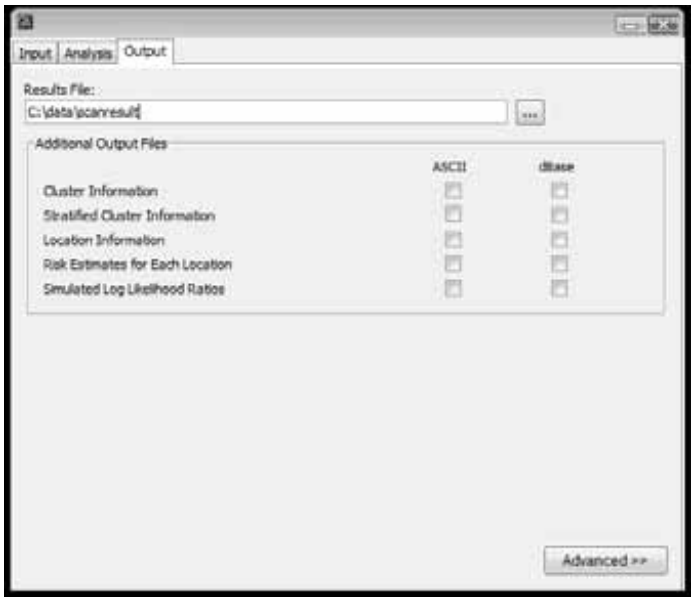
- Download từ http://www.satscan.org/.
- Install
- Khởi động

Nhập dữ liệu trong **SaTScan**




Input
Analysis
Output
Type of Analysis
Retrospective Analyses:
Purely Spatial
Purely Temporal
Space-Time
Prospective Analyses:
Purely Temporal
Space-Time
Probability Model
Discrete Scan Statistics:
Poisson
Bernoulli
Space-Time Permutation
Multinomial
Ordinal
Exponential
Normal
Continuous Scan Statistics:
Poisson
Scan For Areas With:
High Rates
Low Rates
High or Low Rates
Time Aggregation
Units: Year
Month
Day
Length: 1 Years
Monte Carlo Replications (0, 9, 999, or value ending in 999): 999
Advanced >>


Advanced Analysis Features
Spatial Window
Temporal Window
Space and Time Adjustments
Inference
Maximum Spatial Cluster Size
20.0 percent of the population at risk (<= 50%, default = 50%)
50.0 percent of the population defined in the max circle size file (<= 50%)
is a circle with a 1.0 Cartesian units radius
Include Purely Temporal Clusters (Spatial Size = 100%)
Spatial Window Shape
Circular
Elliptic
Non-compactness penalty: Medium
Use Isotonic Spatial Scan Statistic
Set Defaults
Close




```
                    SaTScan v8.0.1


Program run on: Tue Aug 25 15:28:16 2009

Purely Spatial analysis
scanning for clusters with high rates
using the Bernoulli model.
_____________________________________________________________

SUMMARY OF DATA

Study period....................: 1997/1/1 - 1997/12/31
Number of locations.............: 482
Total population................: 482
Total number of cases..........: 186
_____________________________________________________________
```

```
SECONDARY CLUSTERS

2.Location IDs included.: 199, 220, 149, 215, 232, 247, 210,
                          236, 146, 147, 244
  Coordinates / radius..: (23.3591,22.0043) / 1.71
  Population.............: 11
  Number of cases........: 11
  Expected cases.........: 4.24
  Observed / expected...: 2.59
  Relative risk..........: 2.69
  Log likelihood ratio..: 10.679565
  Monte Carlo rank......: 11/1000
  P-value................: 0.011
```

```
MOST LIKELY CLUSTER

1.Location IDs included.: 237, 266, 174, 185, 250, 222, 56, 81,
                          258, 243, 228, 201, 182, 184, 204,
                          214, 275, 203, 197, 85, 193, 191, 251,
                          196, 180, 280, 238, 192, 46, 200, 190,
                          194, 41, 187, 261, 245, 265, 202, 40,
                          260, 216, 235, 212, 284, 268, 79, 480,
                          69, 256, 286, 227, 104, 73, 205, 176,
                          211, 270, 213, 65, 230, 262, 255, 246,
                          273, 84, 209, 177, 75, 60, 248, 264,
                          189, 42, 281, 88, 239, 282
  Coordinates / radius..: (20.0637,16.6265) / 2.60
  Population.............: 77
  Number of cases........: 56
  Expected cases.........: 29.71
  Observed / expected...: 1.88
  Relative risk..........: 2.27
  Log likelihood ratio..: 22.132700
  Monte Carlo rank......: 1/1000
  P-value................: 0.001

SECONDARY CLUSTERS

2.Location IDs included.: 199, 220, 149, 215, 232, 247, 210,
```

Tạo bản đồ có trung tâm dịch bằng R sau khi đã có kết quả từ SaTScan

```
par(pty="s")
polymap(ak.poly, xlab="Easting", ylab="Northing")
points(prv.ctl$X,prv.ctl$Y, pch=1, cex=0.7, col=1)
points(prv.cas$X,prv.cas$Y, pch=16 , cex=0.7, col=4)
cluz <- seq(0, 2*pi, length=1000)
clux <- sin(cluz)
cluy <- cos(cluz)
polygon((clux*2.60)+20.0637, (cluy*2.60)+16.6265, col=4, dens=0,
    lwd=3)
polygon((clux*1.71)+23.3591, (cluy*1.71)+22.0043, col=4, dens=0,
    lwd=3)
```