Metodologia epidemiologica:

Analisi epidemiologica con STATA

Alessandro Marcon

Unit of Epidemiology and Medical Statistics, DPHCM, University of Verona, Italy

(alessandro.marcon@univr.it)

March-May 2015

Course outline

1. Basics with STATA
   - Getting started with STATA
   - Data management
   - Descriptive statistics
   - Basic inference
2. Cross-sectional studies
3. Case-control studies
4. Cohort studies (with person-time data)
5. Standardization
6. Sensitivity, specificity, ROC curves

Materials

- **Datasets** (http://www.dbt.univr.it/?ent=oi&aa=2014%2F2015&codiceCs=S74&codins=4S000647&cs=489&discr=&discrCd=)
- **STATA documentation**
- Online material (examples)
  - http://www.ctspedia.org/do/view/CTSpedia/CourseMaterials
  - http://medicine.utah.edu/ccts/sdbc/files/textbook/Biostats%20and%20Epi%20with%20Stata%20everything%20in%20one%20pdf.pdf
  - http://www.princeton.edu/~otorres/StataTutorial.pdf
- What statistical analysis should I use?
  - http://www.ats.ucla.edu/stat/stata/whatstat/default.htm
  - http://www.graphpad.com/support/faqid/1790/
- http://www.stata.com/

Getting started with STATA

- What do I see on the screen
- How to ask STATA to do things (interactive mode, batch mode, drop-down menus)
- Update STATA
- **How to ask STATA for help**
- How to save STATA output (log file)

Example 1 (children.dta)

- Cross-sectional survey on a hypothetical sample of 648 children
- Data on 15 variables:
  - wheezing, asthma, rhinitis (health outcomes)
  - id, sex, age, height, weight, birthweight (child's characteristics)
  - marker, diagtest (blood biomarker data)
  - foreparent, foreign, peducation, traffic (data on family/house)

Data management - 1

- Cleaning memory
- Loading data in STATA (inserting, importing, opening)
- The data matrix
- **Always get a snapshot of your data as a first thing!** (desc, codebook, count, list, sum, mdesc)

Descriptive analysis

- Basic graphs (histogram, boxplot, scatterplot)
- Frequency distributions (n, %)
- Descriptive statistics (position, dispersion, symmetry, kurtosis)

Wake up! Exercises (1)

1. Does «marker» have a symmetrical distribution?
2. Compare the distribution of «marker» between genders
3. Which variable between «weight» and «birthweight» show greater variability?

Data management - 2

- Setting conditions (in, if, by)
- Sorting and combining datasets
- Managing variables (generating, renaming, recoding, labeling, dropping)
- Saving data

Exercises (2)

1. I tried to compute the 97th percentile of age: **tabstat age, stat(p97)**…this doesn't work! What's the mistake?
2. Create a new categorical variable which classifies subjects in your dataset according to the tertiles of "age"(i.e. three equally-sized groups)
3. Use this new variable to describe the association between age and wheezing

Basic inference

- Comparing means of 2 independent samples
- Comparing proportions of 2 independent samples
- Calculating confidence intervals
- Measuring the association between 2 quantitative variables

Exercises (3)

1. Is there an association between current weight and birth weight?
2. Is birth weight similar for children born in Italy and children born abroad?
3. Compute 99% confidence intervals for the prevalence of "wheezing" for males and females