Introduksjon til STATA

Pål Romundstad

Introduksjon til STATA

HVORFOR?

STATA vs SPSS

SPSS Fordeler:

- Generelt lettere å bruke- lavt brukergrensesnitt
- Enkelt å produsere tabeller og grafer
- Det aller meste dekkes av SPSS, men for epidemiologiske analyser og medisinsk statistikk er det klare mangler

STATA fordeler:

- **Rettet mot medisinsk statistikk / epidemiologi**
- Mange flere prosedyrer enn SPSS
- Bedre dokumentasjon
- Rask
- Raskere utvikling enn SPSS
- **Oppdateres over nettet**
- **Nedlasting av brukergenererte program** (søk online via stata)
- Større påvirkningsmuligheter i analysene
- Mata- matrise orientert programmering
- Lav pris

- **Styrkeberegninger**
- **Kalkulator**
- **Immediate commands - analyser på aggregert nivå**

STATA:

- **Diagnostisk testing**
- **Standardiserte rater**
- **Overlevelse/Cox bedre og flere prosedyrer, parametriske modeller for overlevelse**
- **Maximum likelihood estimatorer**
- Telle prosedyrer (poisson, negative binomial ++
- **Clusteranalyser**
  - Korrelerte utkom
    (eks: flernivå, repeterte målinger)
  - Robust SE
    Huber-White correction for heteroskedascity
  - Random effects models/GEE
- **Missing data analyser med multippel imputering (ICE)**
- **Vekting (pweights vs. aweights and iweights)**
- **Likelihood ratio**
- **Survey analyser** mulig å estimere basert på komplekse survey design
- Omfattende ANOVA/GLM rutiner
- Gode rutiner for longitudinale panel data
- **Brukergenererte prosedyrer på nettet**

Hvordan lære STATA

Diverse kurs (Norge og utland)

På egenhånd

online:

- Karolinska: http://www.cpc.unc.edu/services/computer/presentations/statatutorial
- UCLA: http://www.ats.ucla.edu/stat/stata
- Statas hjemmeside med link til andre: http://www.stata.com/links/resources1.html
- ++++

Bøker: se Statas hjemmeside

Hvordan lære STATA

- Mange eksempler i dag er hentet fra kurs i Stata gitt av Nicola Orsini

Karolinska Institutet, Stocholm:

- http://nicolaorsini.altervista.org/sc/ntnu/is.htm

Hensikten med denne introduksjonen

- Bli kjent med hvordan Stata fungerer
- Bli i stand til å generere enkel beskrivende statistikk
- Lære å lage nye og endre variabler
- Kunne lage krysstabeller
- Gjøre enkle statistiske tester og lineær regresjon
- Vise noen eksempler på grafer

Snakker STATA med andre program

- SPSS kan generere/lese STATA filer
- Stattransfer er et program som lett konverterer mellom aktuelle programpakker inkludert SAS, SPSS, Stata, Access, S+, R, Excel…

Ting å være klar over

- Som SPSS er Stata ”record” orientert
- Både vindusmeny og kommandoorientert
- Missing defineres numerisk med ”**.**” og ses av stata som et høyt tall. Missing string som ” ”
  - varname>10 blir ”true” for missing
- Stata er case sensitiv (skiller mellom store og små bokstaver)

Ting å være klar over

- Stata bruker = og ==
  - Ett likhetstegn brukes ved tilordning, to betyr lik ”equals”
- Du kan bruke forkortelser for de fleste kommandoene
  - EKS: generate = ge, di = display, tab= tabulate……..
- Ved større filer må minne settes høyere enn default-

  Eks1: set memory *500000*

  Eks2: set mem 20m, permanently

Tids/datoformat:

Ses numerisk i forhold til 01.01.1960

- -1=31.12.1959
- 1=02.01.1960

- Dato format: %d

Eksempel:

- Display %d 17177
- gir datoen 11jan1977

Stata vinduer

- Command line – her legges kommandoer inn
- Results –resultatet av kommandoene (output)
- Variables – oversikt over variabler i minnet
- Review – oversikt over tidligere brukte kommandoer

Oversikt over fila: browse eller edit

filtyper

- .dta fil: datasettet (tilsvarer .sav i spss)
- .do fil: kommandofil for å gjøre flere kommandoer på rad (tilsvarer syntax- ".sps" i spss)
- .ado fil: kommandofil i STATA
- .gph fil: graph fil
- .smcl fil: Log fil
- .hlp fil: hjelpefil som forklarer rutiner og kommandoer

Hjelp i Stata

• On-line

Hvis du kjenner kommandonavnet

help *kommandonavn*

Hvis du ikke kjenner kommandonavnet

findit *keywords*

search *keywords*

FAQ (Frequently Asked Questions)

- http://ww.stata.com/support/faqs
- UCLA (Resources to help you learn and use Stata)
- http://www.ats.ucla.edu/stat/stata/

Manualer og bøker

Low Birth Weight Data (lowbwt.dta)

Var.	Beskrivelse	Verdier	Navn
1	Identification Code	ID Number	**id**
2	Low Birth Weight	1 bwt<=2500 g, 0 bwt>2500 g	**low**
3	Age of Mother	Years	**age**
4	Weight of Mother at Last Menstrual	Pounds	**lwt**
5	Race	1 = White, 2=Black, 3=Other	**race**
6	Smoking Status During Pregnancy	0 = No, 1= Yes	**smoke**
7	History of Premature Labor	0 = None, 1=One, 2=Two, etc.	**ptl**
8	History of Hypertension	0 = No, 1 = Yes	**ht**
9	Presence of Uterine Irritability	0 = No, 1 = Yes	**ui**
10	Physician Visits First Trimester	0 = None, 1=One 2=Two, etc.	**ftv**
11	Birth Weight	Grams	**bwt**

Source and abstract of the data

Source

Hosmer and Lemeshow (2000) Applied Logistic Regression: Second Edition. These data are copyrighted by John Wiley & Sons Inc. and must be acknowledged and used accordingly. Data were collected at Baystate Medical Center, Springfield, Massachusetts.

Descriptive abstract

Identify risk factors associated withgiving birth to a low birth weight baby (weighing less than 2500 grams).
Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies.

Hente inn datasett i Stata

Hente datasett (extension .dta) til arbeidsminne

use *filename* [, **clear**]

clear option erases all data currently in memory and proceeds with loading the new data from the disk or from a web server.

/correctly specify the location of the file

use "c:\is\lowbwt.dta", clear
use http://nicolaorsini.altervista.org/data/lowbwt.dta

Alternativ: bruk vindusmeny til å lete frem fil

Estimeringskommandoer

Commando varname(s) if…in…using…, options

Eks:
Tabulate wheese smoker if age>30, col

Analysekommandoer

Generelt:

[prefix:] [command] [varlist] [if] [in] [weight[[,options]

- Hakeparantes angir valgfri input
- **if** begrenser analysen til en undergruppe der if statementet er true
- ***options*** angir spesielle ønsker i output (modify the default)

Eks:
tabulate wheese smoker if age>30, col
regress fødselsvekt maternalage if sex==1
logistic sga maternalage if prematur==0, or

Output

- Finnes brukergenererte programmer som lager fine artikkelferdige output
- Kan kopieres direkte til word (bruk tegnsett 8 pkts courier)
- Grafer kan kombineres og kopieres på flere måter

Se på data

- **list** creates a list of values of specified variables and observations (as alternative to **browse**)
- **describe** provides information on the size of the dataset and the names, labels and types of variables.
- **codebook** summarizes a variable in a format designed for printing a codebook (missing value; unique values; descriptive statistics).

Run a command in a subset of the data

• List the variables and observations if age is greater than 30 years
list if age > 30

• List the variables id, low, and race, if race is equal to 1(White women)
list id low race if race == 1

• List all the variables for the first 10 observations of the data
list in 1/10

Save a dataset

- To save the dataset available in the current memory type

save *filename* [, **replace**]

- the **replace** option indicates that if *filename* already exists,Stata will overwrite it

// Example

save lowbwt2.dta, replace

Summary statistics

summarize calculates and displays a variety of summary statistics. If no *varlist* is specified, then summary statistics are calculated for all the variables in the dataset

summarize [*varlist*] [*if*] [*in*] [, **detail**]

where

detail produces additional statistics, including skewness, kurtosis,various percentiles and the four smallest and largest values.

Example of summarize

```
. su bwt, detail
```

Histogram

A histogram is a graphical method for displaying the shape of a distribution

The height of each bar corresponds to its class frequency.

The command **histogram** allows you to control any aspect of the graph (y-axis and x-axis).

Eks:

histogram bwt, frequency

histogram bwt, frequency width(500) start(500) ylabel(0(5)50) xlabel(500(500)5000) normal

histogram bwt, percent width(500) start(500) ylabel(0(3)24, angle(h)) xlabel(500(500)5000) normal addlabel

Box-plot

A box plot provides a visual summary of many important aspects of a distribution

- The box stretches from the lower hinge (defined as the 25th percentile) to the upper hinge (the 75th percentile) and therefore contains the middle half of the values
- The median is shown as a line across the box.

graph hbox bwt

graph hbox bwt , over(smoke)

Table of counts

tabulate produces one- and two-way tables of frequency counts along with various measures of association (**help tabulate**).

// One-way
tabulate *varname* , options

// Two-way
tabulate *varname1 varname2* , options

It has several options:
chi2 to calculate the Pearson-Chi Square test
col to display column percentages
row to display row percentages
+++

Eks: tabulate race

Kakediagram

graph pie , over(race) ///
plabel(_all percent , color(white))

/// betyr at stata skal fortsette å lese på neste linje

```
tabulate low smoke, col
```

```
graph pie , over(low) by(smoke) ///
plabel(_all percent , color(white))
```

```
graph bar (mean) low , over(smoke) ///
blabel(bar, format(%3.2f)) ///
ytitle(Risk low birth weight)
```

The Chi-square test

Let's compare the proportion of low birth weight in the population of smokers and non-smokers women.

We test the null hypothesis that the two proportions are identical.

```
tabulate low smoke, chi2
tabulate low smoke, chi2 expected
```

Without the full data available

- **tabi** is an immediate commands that displays r x c table using the values specified.
- **tabi** #11 #12 [...] \ #21 #22 [...] [\ ...]

  Immediate commands are useful when full data are not available, like in a review of published analysis or sensitivity analysis of observed data

- It has all the options of **tabulate**

tabi 86 44 \ 29 30, chi2 col

Without the full data available

- **tabi** is an immediate commands that displays r x c table using the values specified.
- **tabi** #11 #12 [...] \ #21 #22 [...] [\ ...]

  Immediate commands are useful when full data are not available, like in a review of published analysis or sensitivity analysis of observed data

- It has all the options of **tabulate**

tabi 86 44 \ 29 30, chi2 col

5 misklassifisert

Without the full data available

- **tabi** is an immediate commands that displays r x c table using the values specified.
- **tabi** #11 #12 [...] \ #21 #22 [...] [\ ...]
  Immediate commands are useful when full data are not available, like in a review of published analysis or sensitivity analysis of observed data

- It has all the options of **tabulate**

tabi 86 44 \ 29 30, chi2 col

5 misklassifisert

tabi 86 44 \ 24 35, chi2 col

Tables of summary statistics

tabstat displays summary statistics for a series of numeric variables in a single tables (**help tabstat**)
tabstat *varlist* [, statistics(*statname*) by(*varname*)]
where ***statname*** are descriptive statistics (mean, min, max,sd, var…).

Without the **by()** option is a useful alternative to summarize because it allows you to specify the list of statistics

tabstat bwt , by(race) stat(mean sd n)

Create a new variable-generate

The command **generate** creates a new variable.
The syntaxis:
generate *newvar* = ***exp*** [***if***] [***in***]

where ***exp*** could be any reasonable combination of variables, numbers, operators (help operators) and functions (help functions)

Operators

1. Arithmetic: + (addition); - (subtraction); / (division); * (multiplication); ^ (raise to a power) and prefix ! (negation).

Any arithmetic operation on missing values or an impossible arithmetic operation yields a missing value.

2. String: + (concatenation of two strings), i.e. “Name” + “Surname”.

3. Relational: > (greater than); < (less than); >= (greater than or equal); <= (less than or equal); == (equal); != (not equal)

4. Logical: **&** (and); **|** (or); **!** (not)

Some examples of generate

*** generate birth weight in kilograms**
gen bwtkg = bwt/1000

*** generate the log of birth weight**
gen logbwt = log(bwt)

*** generate body mass index**
gen bmi = weight_kg/height_mt^2

Replacing values

The command **replace** changes values of a variable that already exists

replace *varname* = *exp* [*if*] [*in*]

* **Identify white women with 0 and non-white women with 1**
replace white **= 0 if race == 1**
replace white **= 1 if race == 2 | race == 3**

How to categorize a variable

The most general approach is to copy or clone the variable you want to categorize

gen white = race

and run one or more replacements to get what you want

replace white = 0 if race == 1
replace white = 1 if race == 2 | race == 3

tabulate race2

Select variables and observations

- **drop** eliminates variables that are explicitly listed
- **keep** keeps variables that are explicitly listed (opposite of **drop**)

drop ***varlist***
drop **if** ***exp***

keep ***varlist***
keep **if** ***exp***

Example

*** eliminate the variables ptl and ht**
drop ptl ht

*** keep 3 variables in memory: id, low, and age**
keep id low age

*** keep only women who smoke**
keep if smoke == 1

Sort observations

sort *varlist*
arranges the observations of the current data into ascending order based on the values of the variables in *varlist*
gsort [+|-] *varname*
arranges observations to be in ascending [+] or descending [-] order of the specified variables

// Examples
sort age
gsort –age

Prefix by

- **by *varlist*: *stata_kommando***
- **bysort *varlist*: *stata_kommando***
- repeats the command for each group of observations for which
- the values of the variables in ***varlist*** are the same.
- The prefix **bysort** gets the **sort** and **by** commands in one line

*** for each level of race tabulate low and smoke**

bysort race : table low smoke

Stata som kalkulator

display *exp*
displays strings and values of scalar expressions.

// Example
display sqrt(4)+ 2*log(1)
display exp(0.5)
display "The square root of 4 is " sqrt(4)

Confidence interval for a proportion

Calculation of an exact 95% CI for the proportion of women who smoke

ci smoke, binomial

Immediate form: cii

cii 189 74

cii 189 74, agresti

Confidence interval of the mean

Calculation a 95% CI of the mean birth weight

ci bwt

Let's calculate a 95% CI of the mean birth weight among smoking and non-smoking women

bysort smoke: ci bwt

Immediate form of **ci**

Let's see how to calculate a 95% CI of the mean birth weight if you know only the number of observations, mean, and standard deviation of the sample.

cii 189 2945 729

Comparisons of two independent samples

Let's see how to test the null hypothesis that the mean birth weights in the population of smokers and non-smokers women are identical

The underlying populations are independent (unpaired) and normally distributed.

Is it likely that the observed difference in sample means – 3054 vs 2773 – is the result of chance variation?, Or should we conclude that the discrepancy is due to true difference in populations means?

Two sample t-test

ttest bwt , by(smoke)

ttest bwt , by(smoke) unequal

Regression model for continuous outcome

To investigate the relationship between a continuous outcome variable and a continuous covariate or predictor we can combine the use of:

- two-way graphs
- simple correlation and linear regression

For instance, investigate the association between birth weight (bwt) and mother's weight (lwt)

Twoway graph

twoway *plot* [if] [in] [, *twoway_options*]

where *plot* is defined

(*plottype* varlist , options**)**

and where *plottype* (**help twoway**) can be, **among many others**

plottype	**description**
scatter	scatterplot
line	line plot
area	line plot with shading
lowess	LOWESS line plot
lfit	linear prediction plot
qfit	quadratic prediction plot
bar	bar plot

Scatter plot

scatter is the most powerful graph command (help scatter)

Can be modified through options and the dialog box can help specifying options (db scatter).

The most common options are:

msymbol()	to specify the shape of marker
mcolor()	to specify the color of marker
ylabel()	to specify labels on y-axis
xlabel()	to specify labels on x-axis
ytitle()	to specify a title on y-axis
xtitle()	to specify a title on x-axis
scheme()	to specify the overall look of the graph
legend()	to specify the legend of the graph

twoway scatter bwt lwt

Add some options

```
twoway scatter bwt lwt, ///
ytitle("Birth weight, grams") ///
xtitle("Mother's weight, pounds") ///
ylabel(500(500)5000, angle(horizontal))
///
xlabel(50(25)250) ///
msymbol(o) ///
mcolor(red)
```

Overlay two plots

```
twoway ///
(scatter bwt lwt) ///
(lfit bwt lwt) , ///
ytitle("Birth weight, grams") ///
xtitle("Mother's weight, pounds") ///
ylabel(500(500)5000, angle(h)) ///
xlabel(50(25)250) ///
legend(label(1 "Scatter") label(2 "Trend")) ///
text(4500 210 "r=0.2 p=0.01") ///
scheme(s1mono)
```

Overlay two plots by groups

```
twoway ///
(scatter bwt lwt) ///
(lfit bwt lwt) , ///
ytitle("Birth weight, grams") ///
xtitle("Mother's weight, pounds") ///
ylabel(500(500)5000, angle(h)) ///
xlabel(50(25)250) ///
legend(label(1 "Scatter") label(2 "Trend")) ///
text(4500 210 "r=0.2 p=0.01") ///
scheme(s1mono)  ///
by(smoke)
```

Simple linear regression

```
regress bwt lwt
```

The first variable is the outcome/response followed by the list of covariates or predictors

Regression with categorical covariates

xi:regress bwt i.race

Prefix xi generates dummy variables for the categories
Can be used for a variety of commands and types of regression

To calculate the population mean (95% CI) birth weight among black women

lincom _cons + _Irace_2

Regression models for binary response

Odds ratios, risk ratios, risk differences, and rate ratios can be estimated with generalized linear models (help glm).
Regression models are much more flexible than tables for epidemiologist and allow the construction of multivariable models.
binreg fits generalized linear models for the binomial family.
It estimates odds ratios, risk ratios, and risk differences

binreg low smoke, rr

xi:binreg low smoke i.race, rr

Several options for generation of odds ratios
binreg low smoke, or
logistic low smoke
logit low smoke, or
glm low smoke, family(binomial) eform