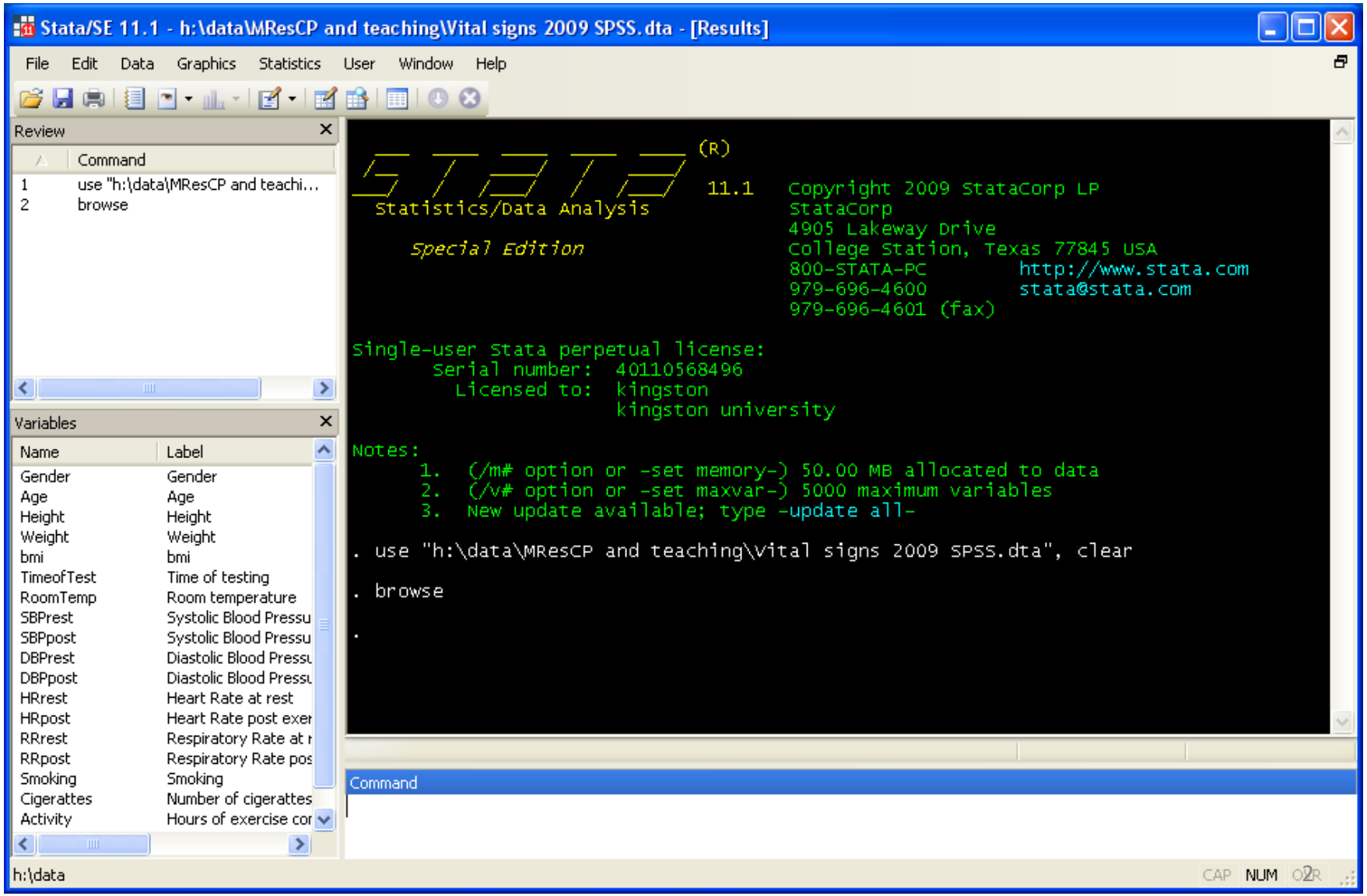
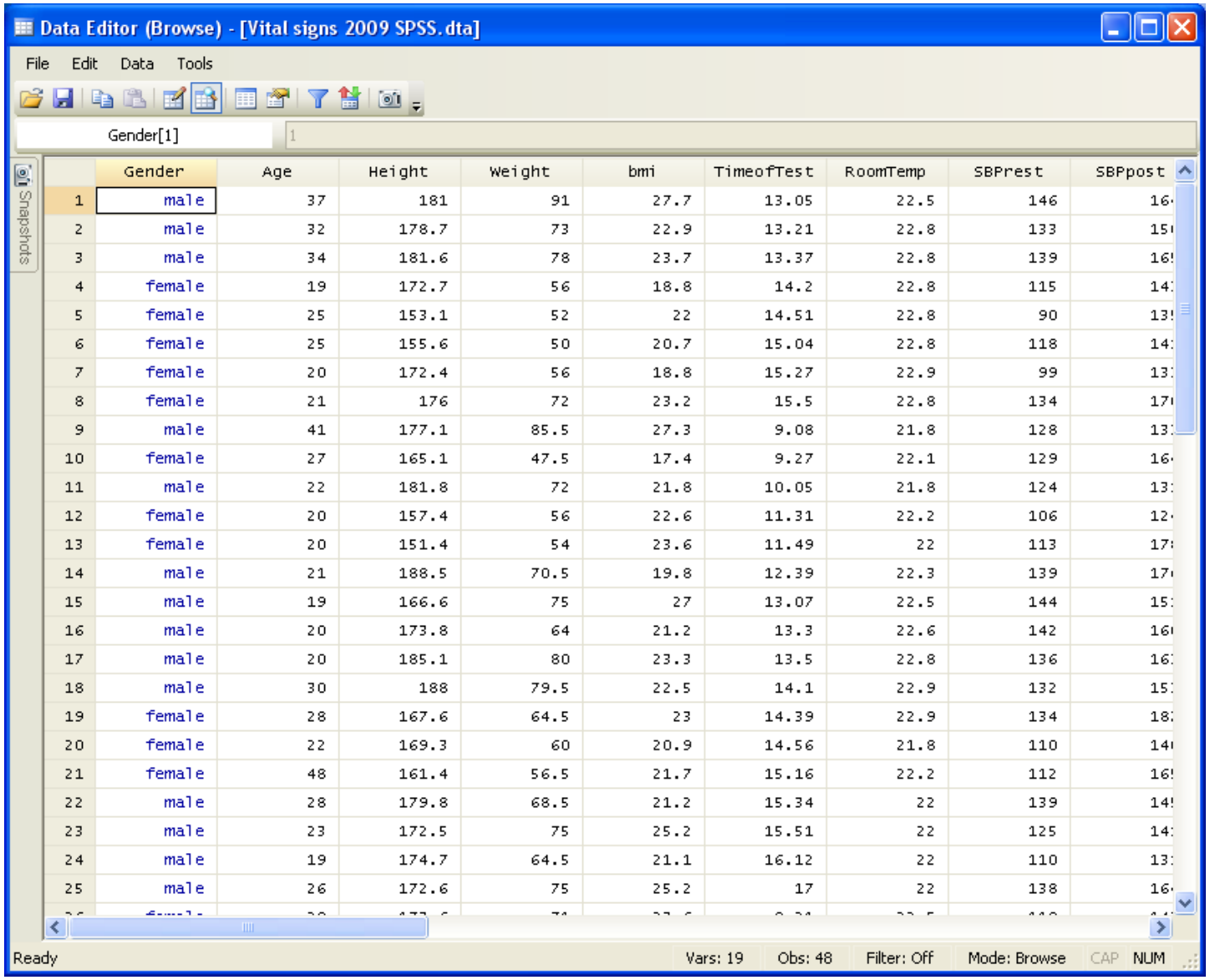
Introduction to Stata

Robert Grant
robert@bayescamp.com
On behalf of Timberlake

Stata

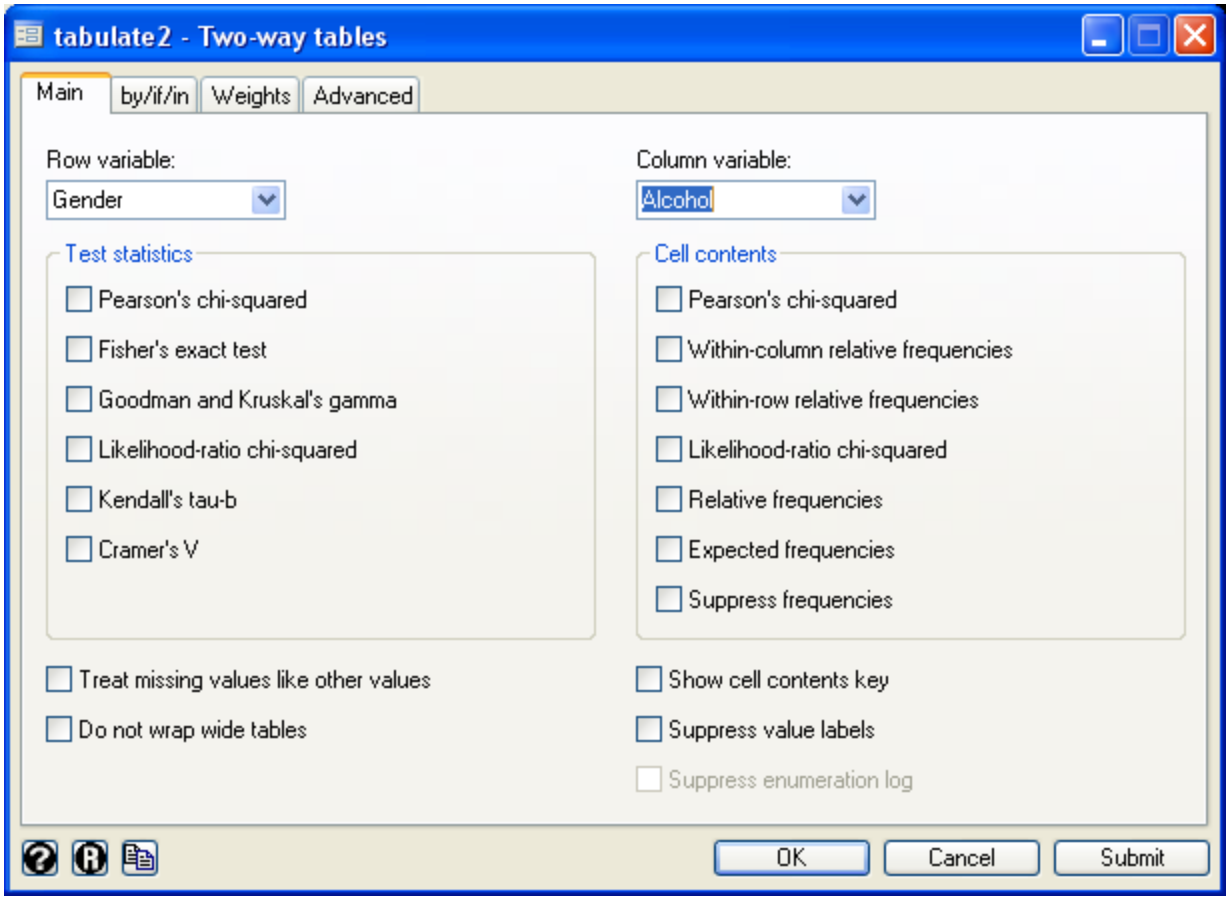

Stata/SE 11.1 - h:\data\MResCP and teaching\Vital signs 2009 SPSS.dta - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command
1 use "h:\data\MResCP and teachi...
2 browse
Variables
Name Label
Gender Gender
Age Age
Height Height
Weight Weight
bmi bmi
TimeofTest Time of testing
RoomTemp Room temperature
SBPrest Systolic Blood Pressu
SBPpost Systolic Blood Pressu
DBPrest Diastolic Blood Pressu
DBPpost Diastolic Blood Pressu
HRrest Heart Rate at rest
HRpost Heart Rate post exer
RRrest Respiratory Rate at r
RRpost Respiratory Rate pos
Smoking Smoking
Cigerattes Number of cigerattes
Activity Hours of exercise cor
(R)
11.1
Statistics/Data Analysis
Special Edition
Copyright 2009 StataCorp LP
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
Single-user Stata perpetual license:
Serial number: 40110568496
Licensed to: kingston
kingston university
Notes:
1. (/m# option or -set memory-) 50.00 MB allocated to data
2. (/v# option or -set maxvar-) 5000 maximum variables
3. New update available; type -update all-
. use "h:\data\MResCP and teaching\Vital signs 2009 SPSS.dta", clear
. browse
.
Command
h:\data
CAP NUM OVR

Stata

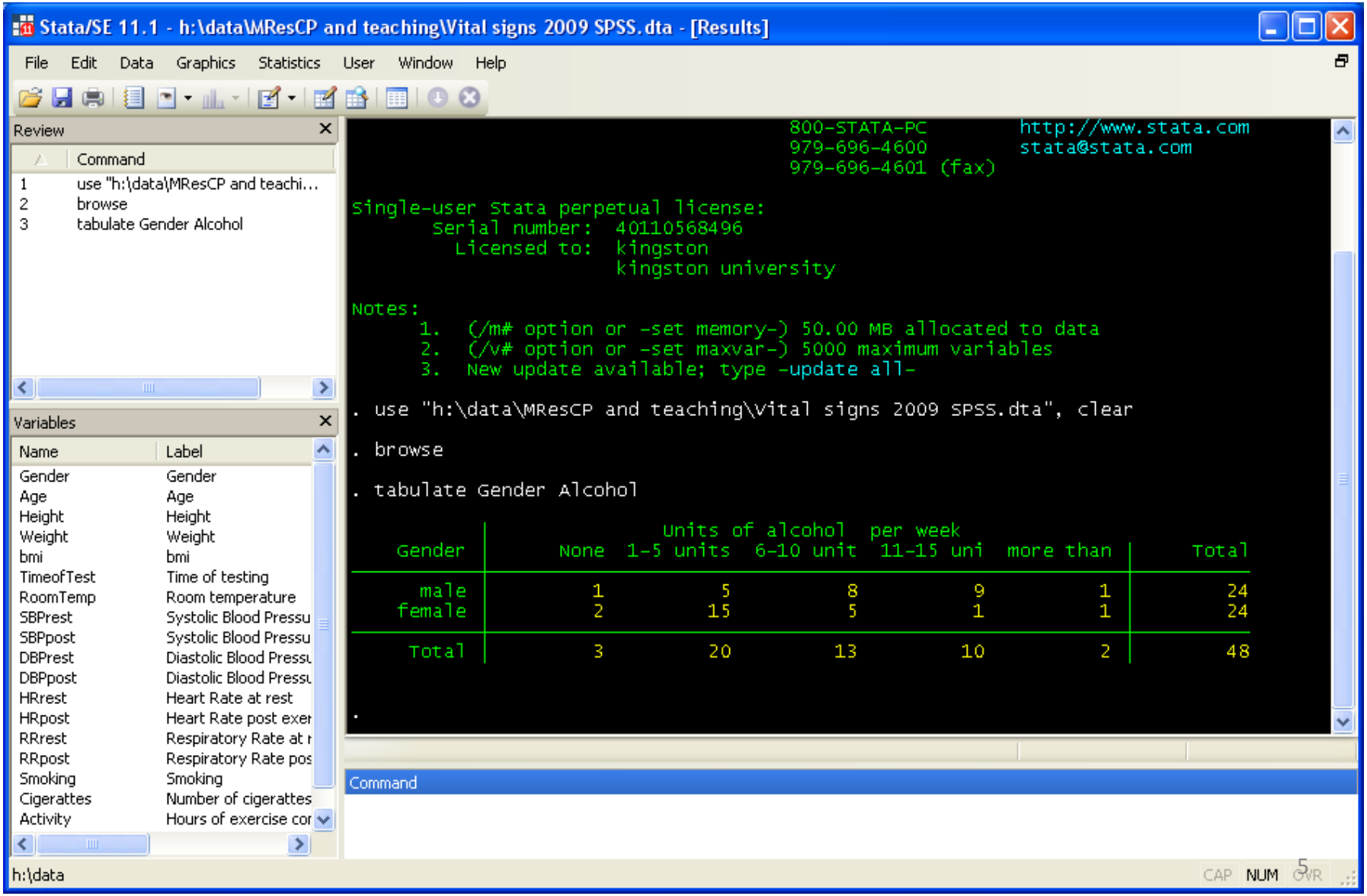


	Gender	Age	Height	Weight	bmi	TimeofTest	RoomTemp	SBPrest	SBPpost
1	male	37	181	91	27.7	13.05	22.5	146	16
2	male	32	178.7	73	22.9	13.21	22.8	133	15
3	male	34	181.6	78	23.7	13.37	22.8	139	16
4	female	19	172.7	56	18.8	14.2	22.8	115	14
5	female	25	153.1	52	22	14.51	22.8	90	13
6	female	25	155.6	50	20.7	15.04	22.8	118	14
7	female	20	172.4	56	18.8	15.27	22.9	99	13
8	female	21	176	72	23.2	15.5	22.8	134	17
9	male	41	177.1	85.5	27.3	9.08	21.8	128	13
10	female	27	165.1	47.5	17.4	9.27	22.1	129	16
11	male	22	181.8	72	21.8	10.05	21.8	124	13
12	female	20	157.4	56	22.6	11.31	22.2	106	12
13	female	20	151.4	54	23.6	11.49	22	113	17
14	male	21	188.5	70.5	19.8	12.39	22.3	139	17
15	male	19	166.6	75	27	13.07	22.5	144	15
16	male	20	173.8	64	21.2	13.3	22.6	142	16
17	male	20	185.1	80	23.3	13.5	22.8	136	16
18	male	30	188	79.5	22.5	14.1	22.9	132	15
19	female	28	167.6	64.5	23	14.39	22.9	134	18
20	female	22	169.3	60	20.9	14.56	21.8	110	14
21	female	48	161.4	56.5	21.7	15.16	22.2	112	16
22	male	28	179.8	68.5	21.2	15.34	22	139	14
23	male	23	172.5	75	25.2	15.51	22	125	14
24	male	19	174.7	64.5	21.1	16.12	22	110	13
25	male	26	172.6	75	25.2	17	22	138	16

Stata

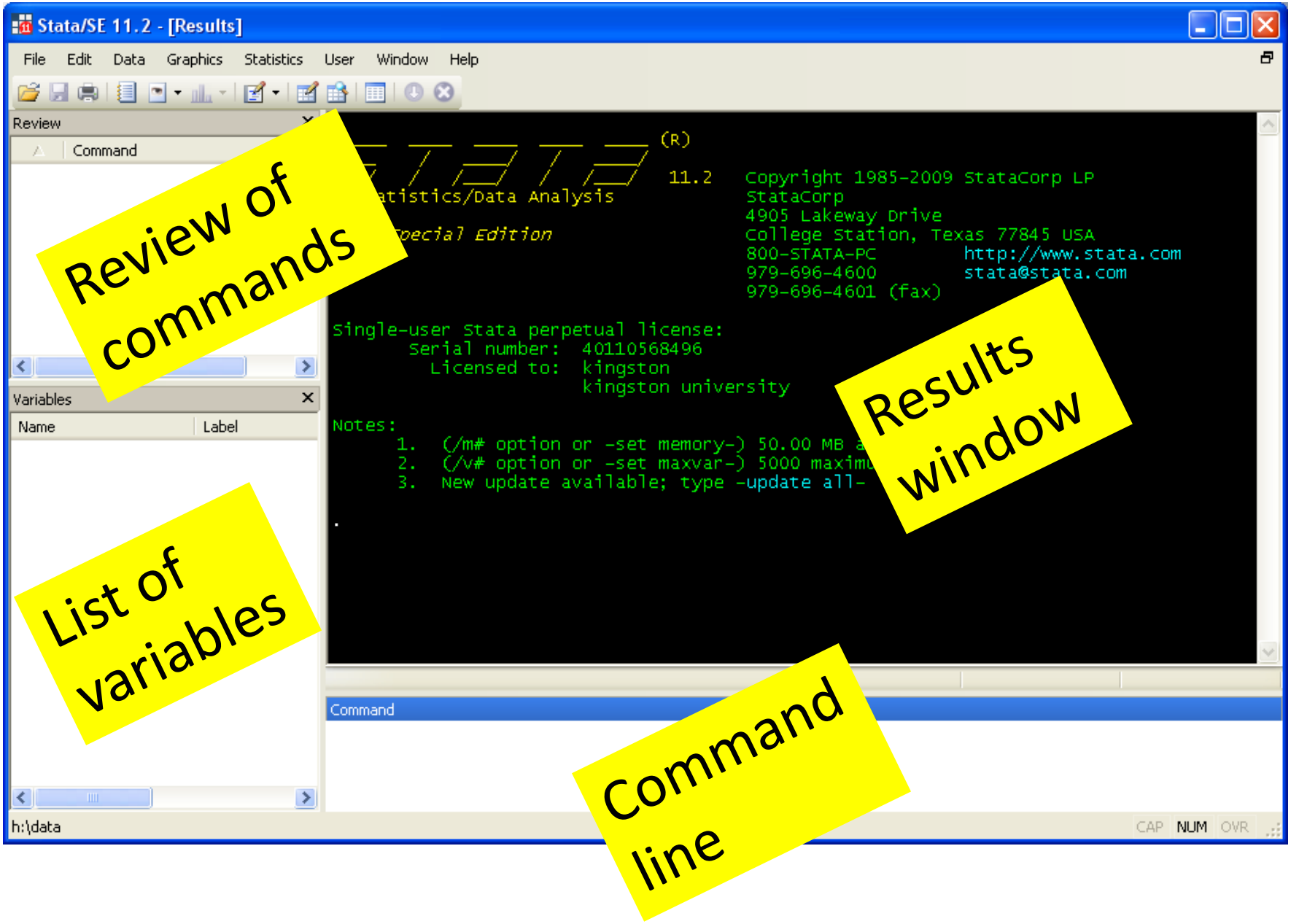

tabulate2 - Two-way tables
Main
by/if/in
Weights
Advanced
Row variable:
Gender
Column variable:
Alcohol
Test statistics
Pearson's chi-squared
Fisher's exact test
Goodman and Kruskal's gamma
Likelihood-ratio chi-squared
Kendall's tau-b
Cramer's V
Cell contents
Pearson's chi-squared
Within-column relative frequencies
Within-row relative frequencies
Likelihood-ratio chi-squared
Relative frequencies
Expected frequencies
Suppress frequencies
Treat missing values like other values
Do not wrap wide tables
Show cell contents key
Suppress value labels
Suppress enumeration log
OK
Cancel
Submit

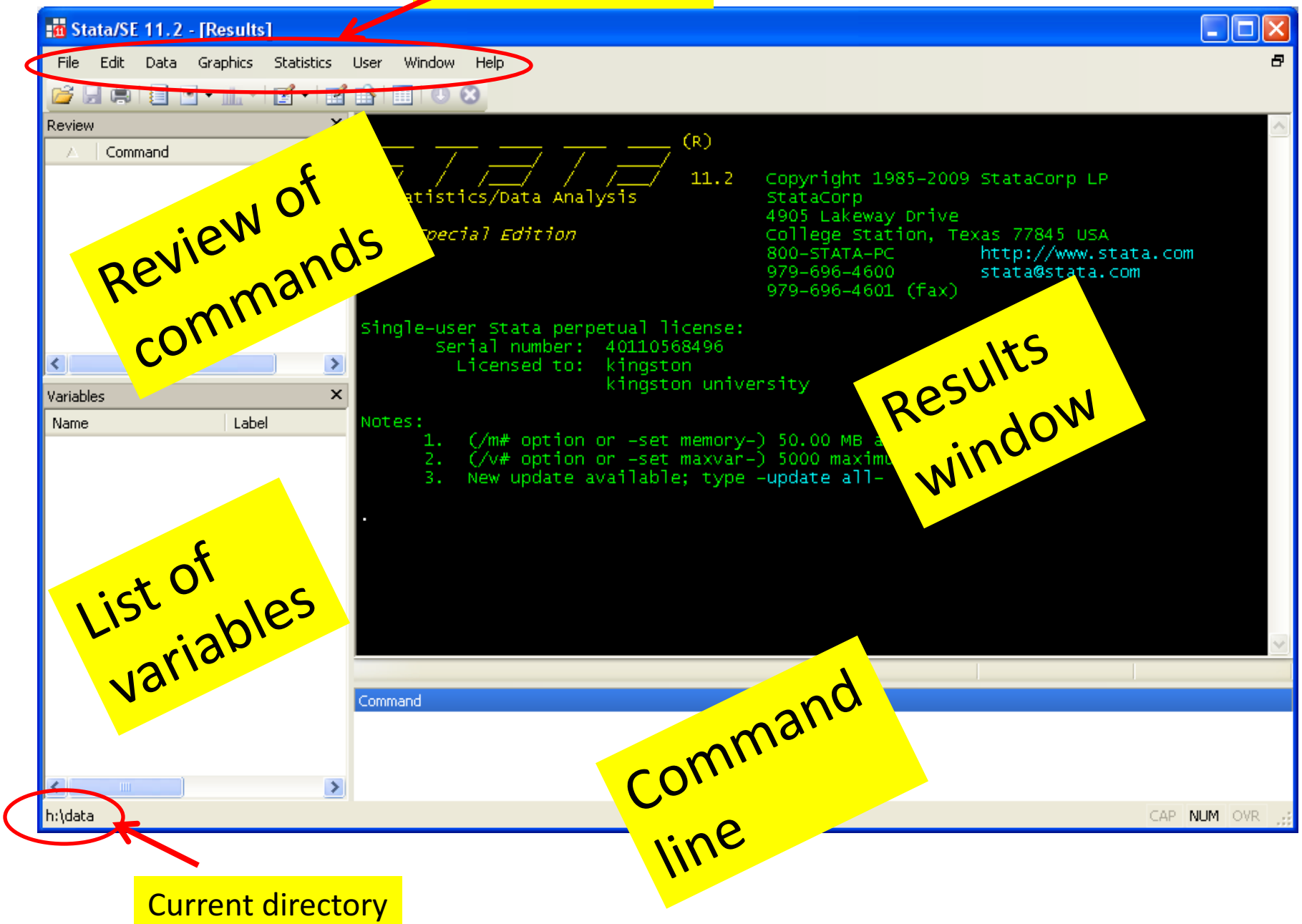
Stata

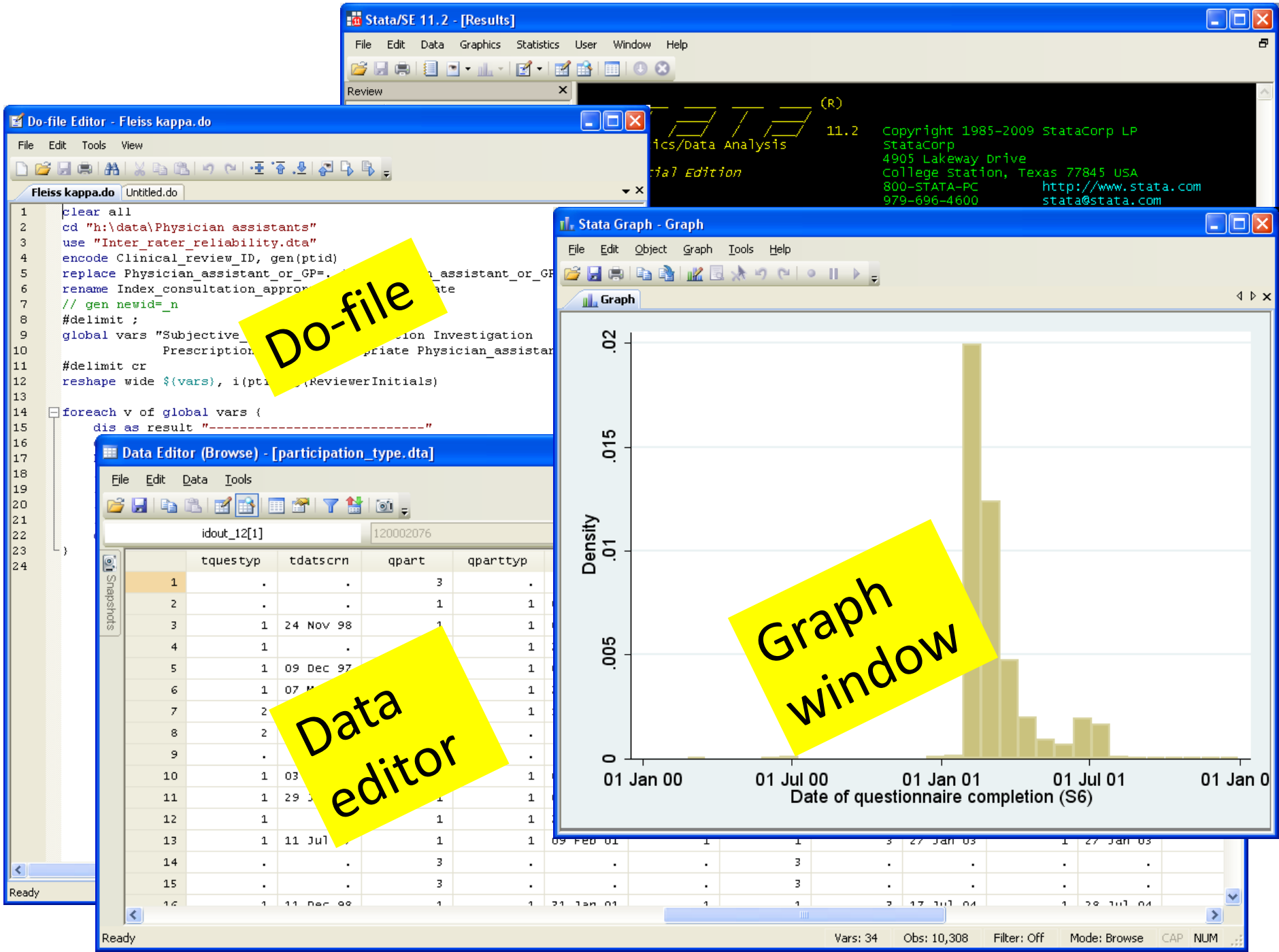

Stata/SE 11.1 - h:\data\MResCP and teaching\Vital signs 2009 SPSS.dta - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command
1 use "h:\data\MResCP and teachi...
2 browse
3 tabulate Gender Alcohol
Variables
Name Label
Gender Gender
Age Age
Height Height
Weight Weight
bmi bmi
TimeofTest Time of testing
RoomTemp Room temperature
SBPrest Systolic Blood Pressu
SBPpost Systolic Blood Pressu
DBPrest Diastolic Blood Pressu
DBPpost Diastolic Blood Pressu
HRrest Heart Rate at rest
HRpost Heart Rate post exer
RRrest Respiratory Rate at r
RRpost Respiratory Rate pos
Smoking Smoking
Cigerattes Number of cigerattes
Activity Hours of exercise cor
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
single-user Stata perpetual license:
Serial number: 40110568496
Licensed to: kingston
kingston university
Notes:
1. (/m# option or -set memory-) 50.00 MB allocated to data
2. (/v# option or -set maxvar-) 5000 maximum variables
3. New update available; type -update all-
. use "h:\data\MResCP and teaching\Vital signs 2009 SPSS.dta", clear
. browse
. tabulate Gender Alcohol
Gender | Units of alcohol per week
None 1-5 units 6-10 unit 11-15 uni more than | Total
male 1 5 8 9 1 | 24
female 2 15 5 1 1 | 24
Total 3 20 13 10 2 | 48
.
Command
h:\data
CAP NUM OVR

Strengths of Stata

- Users can write and share new functions that extend Stata's capability
- Good customer support
- High-quality, flexible graphics
- Easy to get started and learn
- Easy to program
- Cheaper than SPSS or SAS!


Stata/SE 11.2 - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command
Variables
Name Label
(R)
11.2
Copyright 1985-2009 StataCorp LP
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
Single-user Stata perpetual license:
Serial number: 40110568496
Licensed to: kingston
kingston university
Notes:
1. (/m# option or -set memory-) 50.00 MB a
2. (/v# option or -set maxvar-) 5000 maximu
3. New update available; type -update all-
Command
h:\data
CAP NUM OVR
Review of commands
Results window
List of variables
Command line


Drop-down menus
Stata/SE 11.2 - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command
Review of commands
(R)
11.2
Copyright 1985-2009 StataCorp LP
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
Single-user Stata perpetual license:
Serial number: 40110568496
Licensed to: kingston
kingston university
Notes:
1. (/m# option or -set memory-) 50.00 MB a
2. (/v# option or -set maxvar-) 5000 maximu
3. New update available; type -update all-
Results window
Variables
Name
Label
List of variables
Command
Command line
h:\data
Current directory
CAP NUM OVR

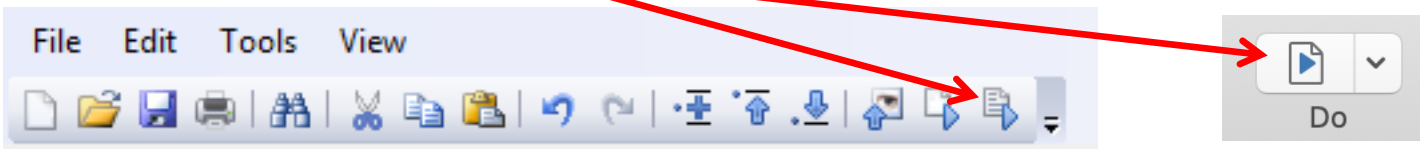

Stata/SE 11.2 - [Results]
File Edit Data Graphics Statistics User Window Help
Review
(R)
11.2
Copyright 1985-2009 StataCorp LP
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
Do-file Editor - Fleiss kappa.do
File Edit Tools View
Fleiss kappa.do
Untitled.do
clear all
cd "h:\data\Physician assistants"
use "Inter_rater_reliability.dta"
encode Clinical_review_ID, gen(ptid)
// gen newid=_n
#delimit ;
#delimit cr
reshape wide ${vars}, i(ptid) j(ReviewerInitials)
foreach v of global vars {
dis as result "----------"
}
Do-file
Data Editor (Browse) - [participation_type.dta]
File Edit Data Tools
idout_12[1]
120002076
Snapshots
tquestyp tdatscrn qpart qparttyp
Data editor
Ready
Vars: 34 Obs: 10,308 Filter: Off Mode: Browse CAP NUM
Stata Graph - Graph
File Edit Object Graph Tools Help
Graph
Density
0
.005
.01
.015
.02
01 Jan 00
01 Jul 00
01 Jan 01
01 Jul 01
Date of questionnaire completion (S6)
Graph window

Some Stata technicalities

- Four “flavors”: Small Stata, IC, SE, MP
- Increasing memory capacity
- Serious research requires SE
- MP is for multi-core processors (CPU)
- Own .dta data file format – but can import

Ways to carry out analyses

- Drop-down menu, select options
- Type command and press <return>
- Keep commands in a do-file, highlight them and click <Do> or press <Control-D> / <Shift-Cmd-D



- Do-files are the best approach

Advantages of using do-files

- Instant audit trail
- Facilitates collaboration
- Helps you remember what you did
- Saves you a lot of time
- Some journals will publish as web supplement
- Reproducible research

Good practice with do-files

- Have one file for the project if possible
- Include all data cleaning and editing
- Add lots of annotations to help you remember and others to understand what you did
- Use **clear** and **replace** options and **capture** to avoid error messages
- Any information that is used repeatedly can be stored in memory once using **local** or **global**

Log files

- log using “mylog.smcl”, smcl replace
- Records everything that happens into a file in Stata’s own SMCL format, until you type:
- log close
- You can also use plain text format
- Logging is a very good idea, even if you only look at it when things go wrong.

Opening data files

- The File-Open menu works like any other software, and looks for Stata .dta files
- You can type **use "mydata.dta"** to open from the command line
- You can import from a variety of other formats
- **import excel** and **import delimited** are import commands for other file formats
- We will use built-in example data with the **sysuse** command

Summarizing data

- These are commands you will use a lot:
- **summarize bp_before**
- **tabulate agegrp**
- **tabulate sex agegrp, row**
- **tabstat bp_after, by(agegrp) stat(n mean sd)**

Finding help

- The single most important skill!
- Even experienced users look stuff up
    - **help tabstat**
    - **h tabstat**
    - **hsearch table statistics**
- Statalist e-mail group and archives
- The Stata Journal
- Stata Press books
- Google "… in Stata"

Processing data

- We will cover this in a lot more detail in the next two lectures
- Some ways to make and change data
- **generate change = bp_before – bp_after**
- **edit** … then click and type
- **recode bp_after (0/139 = 1) (140/999 = 2), generate(bp_group)**

Labelling data

- **label var change “Change in blood pressure”**
- **label define bplab 1 “Final BP < 140” 2 “Final BP 140+”**
- **label values bp_group bplab**
- Your labels are added into tables and graphs

Entering data directly

- **clear all**
- **set obs 100** makes 100 blank rows
- **generate x=.** Makes a variable called x which is empty
- **edit**
- And off you go…

Miscellaneous

- **set more off, permanently** gets rid of the annoying –more– link that you have to click on to see your output.
- Preferences menu to change color scheme
- **cd** changes the directory Stata is working in
- You can abbreviate a lot of commands (e.g. **tab sex bp_group**)
- And variable names too, if unambiguous

Selecting part of the data

- **ttest bp_after=bp_before** is a t-test on all data
- **ttest bp_after=bp_before if patient<101** does the same thing, but only on patients with ID numbers up to 100
- **ttest bp_after=bp_before in 21/40** does a t-test on rows 21 to 40 in the file
- **bysort sex: ttest bp_after=bp_before** will run the t-test once for men and once for women

More tricks

- **quietly:** will suppress output from any command that follows
- Most commands save extra stats that you can retrieve with names like **r(mean)** or **e(b)** – they are listed at the end of help files
- Commands that are very long can ///
- be split over lines using three slashes like this
- Stata uses scientific notation: -4.72e-09 means -0.00000000472