Introduction to Stata

17.871

Spring 2012

The role of statistical packages in research

- Obvious answer
  - Manage data
  - Carry out appropriate statistical tests
  - Assist in displaying data
- Less obvious answer
  - Channel the type of research you are likely to do
    - Limitations as to variables and cases
    - Types of analysis is sometimes guided by choice of package

Analysis -> Packages

- Baby exercises
  - Minitab, spreadsheets
- Time series
  - TSP
- Cross-sectional
  - SPSS, SAS
- Time series & cross-sectional
  - Stata, R

Logic of quant research in this class

$$y_i = f(x_i, \beta, \varepsilon_i)$$

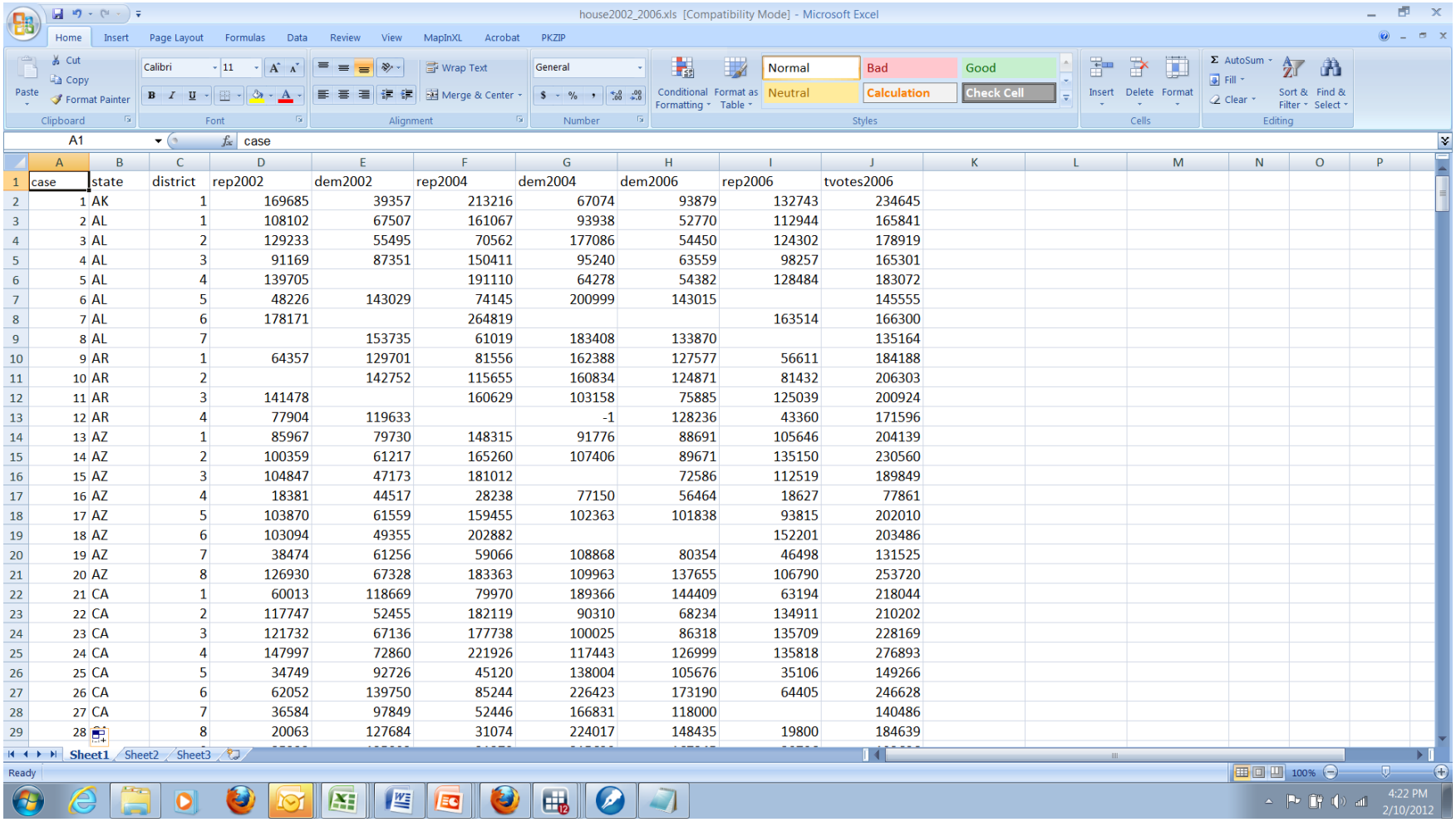
Logic of data setup:

	V_1	V_2	…	V_j
Obs_1				
Obs_2				
…				
Obs_i				

Example, VRS Data

HRHHID	GESTCEN	PES1	PES8
199960521980910	63	2	4
160916068405549	63	2	-3
941159210626002	63	2	6
941159210626002	63	2	6
941159210626002	63	2	6

Example, House Elections



house2002_2006.xls [Compatibility Mode] - Microsoft Excel

case	state	district	rep2002	dem2002	rep2004	dem2004	dem2006	rep2006	tvotes2006
1	AK	1	169685	39357	213216	67074	93879	132743	234645
2	AL	1	108102	67507	161067	93938	52770	112944	165841
3	AL	2	129233	55495	70562	177086	54450	124302	178919
4	AL	3	91169	87351	150411	95240	63559	98257	165301
5	AL	4	139705		191110	64278	54382	128484	183072
6	AL	5	48226	143029	74145	200999	143015		145555
7	AL	6	178171		264819			163514	166300
8	AL	7		153735	61019	183408	133870		135164
9	AR	1	64357	129701	81556	162388	127577	56611	184188
10	AR	2		142752	115655	160834	124871	81432	206303
11	AR	3	141478		160629	103158	75885	125039	200924
12	AR	4	77904	119633		-1	128236	43360	171596
13	AZ	1	85967	79730	148315	91776	88691	105646	204139
14	AZ	2	100359	61217	165260	107406	89671	135150	230560
15	AZ	3	104847	47173	181012		72586	112519	189849
16	AZ	4	18381	44517	28238	77150	56464	18627	77861
17	AZ	5	103870	61559	159455	102363	101838	93815	202010
18	AZ	6	103094	49355	202882			152201	203486
19	AZ	7	38474	61256	59066	108868	80354	46498	131525
20	AZ	8	126930	67328	183363	109963	137655	106790	253720
21	CA	1	60013	118669	79970	189366	144409	63194	218044
22	CA	2	117747	52455	182119	90310	68234	134911	210202
23	CA	3	121732	67136	177738	100025	86318	135709	228169
24	CA	4	147997	72860	221926	117443	126999	135818	276893
25	CA	5	34749	92726	45120	138004	105676	35106	149266
26	CA	6	62052	139750	85244	226423	173190	64405	246628
27	CA	7	36584	97849	52446	166831	118000		140486
28	CA	8	20063	127684	31074	224017	148435	19800	184639

Using Stata to Analyze Data in Matrix Form

- Question: Did Ron Paul do better in Iowa in 2012, compared to 2008 in counties with college students?
- Data sources:
  - 2008: Des Moines Register web site
  - 2012: Iowa Republican Party, Google Doc (https://www.google.com/fusiontables/DataSource?dsrcid=2475248)

Switch over to Stata run-through

Return from Stata run-through

- Why would you use different input commands?

insheet

- Data is output from a spreadsheet into “csv” or “comma-delimited” format
- Data is a simple *I* x *J* matrix, and all the variables are separated either by a tab or comma
- Stata is now smart enough to figure out that the first line of the file contains the variable names

insheet

Assume the following file was created by outputting a file from Excel in csv format:

HRHHID	GESTCEN	PES1	PES8
19996052198091O	63	2	4
160916068405549	63	2	-3
941159210626002	63	2	6
941159210626002	63	2	6
941159210626002	63	2	6

insheet using *filename*

infile

- Data is not in Stata format, is in an ASCII file, but is *not* separated *only* by a tab or comma (e.g., by a space)

insheet

Assume the following file was created using an ASCII text editor (e.g., EMACS), and that spaces separate the variables:

```
19996052198091O 63 2 4
160916068405549 63 2 -3
941159210626002 63 2 6
941159210626002 63 2 6
941159210626002 63 2 6
```

infile HRHHID GESTCEN PES1 PES8 using *filename*

Or

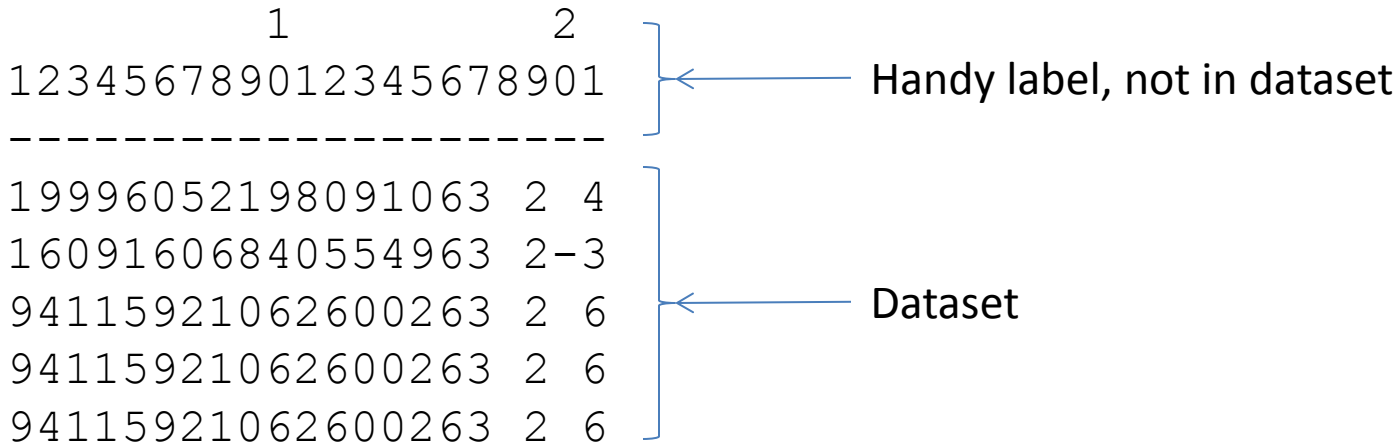
infile str HRHHID GESTCEN PES1 PES8 using *filename*

infix

- Data is in an ASCII file, but you cannot rely on spaces, commas, or other standard "delimeters" to separate variables
- Datasets may have observations on more than one line

infix

Assume the following file was created using an ASCII text editor:



infix HRHHID 1-15 GESTCEN 16-17 PES1 18-19 PES8 20-21 using *filename*

Or

infile str15 HRHHID 1-15 GESTCEN 16-17 PES1 18-19 PES8 20-21 using *filename*

House Roll Call votes in the 27th Cong.

```
01R32703120029000340 1ADAMS      16555561666166111122222626111661196611611611666 6
02R32703120029000340 1ADAMS      66616111611166611666616611116611611619161166666 6
03R32703120029000340 1ADAMS      66116661111661166666119166111661169916111616161 1
04R32703120029000340 1ADAMS      16116661616611916991111661611661166161661661161 1
05R32703120029000340 1ADAMS      16666661611161916616116166666666161111666616111 1
06R32703120029000340 1ADAMS      16666616111616116611111166166666112661166166666 6
07R32703120029000340 1ADAMS      69666161666661116911161111116116661111116161161 6
08R32703120029000340 1ADAMS      11916666666616666661116666699999116166116999916 1
09R32703120029000340 1ADAMS      66661611116111666696616161116611166661666161111 9
10R32703120029000340 1ADAMS      61161666116166161666116116111111111616111991996 6
11R32703120029000340 1ADAMS      11619166616116116696616111616661161169116916666 6
12R32703120029000340 1ADAMS      61116669966161666116616111616611116111166116666 1
13R32703120029000340 1ADAMS      61166611661616166661661696166661166616666166611 1
14R32703120029000340 1ADAMS      11616111116116661161116666166616661661661666116 6
15R32703120029000340 1ADAMS      61161661161611116116111116166111661116611166616 6
16R32703120029000340 1ADAMS      16111661911666661661161616666196666161161661611 1
17R32703120029000340 1ADAMS      66111616111161166616666166661111616161666661111 1
18R32703120029000340 1ADAMS      11166699161666161666111166161661161611611616166 6
19R32703120029000340 1ADAMS      16661661116116116111661116166666611166611191161 1
20R32703120029000340 1ADAMS      61661661611916166616619666611966661166166611111 6
21R32703120029000340 1ADAMS      61111161111161
01R32744980032000911 1ALFORD     65555599661691616555256511116116111911991999 99
02R32744980032000911 1ALFORD     91691666116961166166116199991161161111116116999 9
```

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
01R32703120029000340 1ADAMS      1655556166616611112222262611166119661161161166666
02R32703120029000340 1ADAMS      6661611161116661166661661111661161161916116666666
03R32703120029000340 1ADAMS      6611666111166116666611916611166116699161116161611
04R32703120029000340 1ADAMS      1611666161661191699111166161166116616616616611611
05R32703120029000340 1ADAMS      1666666161116191661611616666666616111166661611111
06R32703120029000340 1ADAMS      1666661611161611661111166166666112661166166666666
07R32703120029000340 1ADAMS      6966616166666111691116111111611666111111616116116
08R32703120029000340 1ADAMS      1191666666661666666111666669999911616611699991161
09R32703120029000340 1ADAMS      6666161111611166669661616111661116666166616116119
10R32703120029000340 1ADAMS      6116166611616616616611611611111111161611199199 66
11R32703120029000340 1ADAMS      1161916661611611666966161116166611611669116916666
12R32703120029000340 1ADAMS      6111666996616166611661611161661111611116611666661
13R32703120029000340 1ADAMS      6116661166161616666166169616666116661666616666611
14R32703120029000340 1ADAMS      1161611111611666116111666616661666166166166616661166
15R32703120029000340 1ADAMS      6116166116161111611611111616611166116611661116666166
16R32703120029000340 1ADAMS      1611166191166666166116161666619666616111616616611
17R32703120029000340 1ADAMS      6611161611116116661666616666111161616166666111111
18R32703120029000340 1ADAMS      1116669916166616166611116616166116161161161616666
19R32703120029000340 1ADAMS      1666166111611611611166111616666661116661119116 11
20R32703120029000340 1ADAMS      6166166161191616661661966661196666116616661111 16
21R32703120029000340 1ADAMS      61111161111161
01R32744980032000911 1ALFORD     6555559966169161655552565111161161119111991999 99
02R32744980032000911 1ALFORD     9169166611696116616611619999116116111111611699 99

VAR #  0004              WIDTH = 0002      MD=0    DK 01    COL 07-08      H27

          STATE:
          ......
          NEW ENGLAND                        BORDER STATES
          ...........                        ............
          01.  CONNECTICUT                   51.  KENTUCKY
          02.  MAINE                         52.  MARYLAND
          03.  MASSACHUSETTS                 53.  OKLAHOMA
```

```
          1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
01R32703120029000340 1ADAMS      1655556166616611112222262611166119661161161166666
02R3270312002900340 1ADAMS      6661611161116661166661661111661161161916116666666
03R3270312002900340 1ADAMS      6611666111166116666611916611166116699161116161611
04R3270312002900340 1ADAMS      1611666161661191699111166161166116616616616611611
05R3270312002900340 1ADAMS      1666666161116191661611616666666616111166661611111
06R3270312002900340 1ADAMS      1666661611161611661111116616666611266116616666666
07R3270312002900340 1ADAMS      6966616166666111691116111111611666111111616116116
08R3270312002900340 1ADAMS      1191666666661666666111666669999911616611699991161
09R3270312002900340 1ADAMS      6666161111611166669661616111661116666166616116119
10R3270312002900340 1ADAMS      6116166611616616616611611611111111161611199199 66
11R3270312002900340 1ADAMS      1161916661611611666966161116166611611669116916666
12R3270312002900340 1ADAMS      6111666996616166611661611161661111611116611666661
13R3270312002900340 1ADAMS      6116661166161616666166169616666116661666616666611
14R3270312002900340 1ADAMS      1161611111611666116111666616661666166166166166611 66
15R3270312002900340 1ADAMS      6116166116161111611611111616611166116611661116666166
16R3270312002900340 1ADAMS      1611166191166666166116161666619666616116166166611
17R3270312002900340 1ADAMS      6611161611116116666166661666611116161616666611111
18R3270312002900340 1ADAMS      1116669916166616166611116616166116161161161616666
19R3270312002900340 1ADAMS      1666166111611611611166111616666661116661119116 11
20R3270312002900340 1ADAMS      6166166161191616661661966661196666116616661111 16
21R3270312002900340 1ADAMS      61111161111161
01R32744980032000911 1ALFORD    6555559966169161655552565111161161119111991999 99
02R3274498003200911 1ALFORD     9169166611696116616611619999116116111111611699 99
```

VAR # 0005 WIDTH = 0002 MD=0 DK 01 COL 09-10 H27

DISTRICT NUMBER:

.....

CODED BLANK FOR SENATE.

AT-LARGE DISTRICTS ARE CODED 98,97,96, ACCORDING TO ALPHABETICAL ORDER OF NAMES OF OCCUPANTS. NO DISTINCTION BETWEEN THE VARIOUS KINDS OF AT-LARGE DISTRICTS IS MADE.

DUE TO REPLACEMENTS WITHIN A CONGRESS, TWO MEMBERS MAY LEGITIMATELY HAVE THE SAME DISTRICT NUMBER WITHIN A STATE.

```
          1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
01R3270312 0029 0003401ADAMS      1655556166166111122222626111661196611611666 6
02R327031200290003401ADAMS        6661611161116661166661661111661161161916116666 66
03R327031200290003401ADAMS        6611666111166116666611916611166116699161116161611
04R327031200290003401ADAMS        1611666161661191699111166161166116616616611611
05R327031200290003401ADAMS        1666666161116191661611616666666616111166661611 11
06R327031200290003401ADAMS        1666661611161611661111166166666112661166166666 6
07R327031200290003401ADAMS        6966616166666111691116111111611666111111616116 16
08R327031200290003401ADAMS        1191666666661666666111666669999911616611699991 61
09R327031200290003401ADAMS        6666161111611166669661616111661116666166616111 19
10R327031200290003401ADAMS        6116166611616616616611611611111111161611199199 66
11R327031200290003401ADAMS        1161916661611611666966161116166611611669116916 66
12R327031200290003401ADAMS        6111666996616166611661611161661111611116611666 61
13R327031200290003401ADAMS        6116661166161616666166169616666116661666616666 11
14R327031200290003401ADAMS        1161611111611666116111666616661666166166166611 66
15R327031200290003401ADAMS        6116166116161111611611111616611166116611661116 66
16R327031200290003401ADAMS        1611166191166666166116161666619666616111616616 11
17R327031200290003401ADAMS        6611161611116116661666616666111161616166666111 11
18R327031200290003401ADAMS        1116669916166616166611116616166116161161161616 66
19R327031200290003401ADAMS        1666166111611611611166111616666661116661119116 11
20R327031200290003401ADAMS        6166166161191616661661966661196666116616661111 16
21R327031200290003401ADAMS        61111161111161
01R3274498 0032 0009111ALFORD     6555559966169161655552565111161161119111991999 99
02R327449800320009111ALFORD       9169166611696116616611619999116116111111611699 99
```

VAR # 0020 SESSION 1 WIDTH = 0001 MD=0 DK 01 COL 42-42 H27

G-10- -27A J 27-1-39 JUNE 7, 1841
H271004 Y=66 N=149 MALLORY, VA.
TO ADJOURN, IN ORDER TO END DEBATE ON THE ADOPTION OF THE HOUSE RULES. ADOPTION OF THE RULES WOULD PREVENT RECEIVING ANY ABOLITION PETITIONS. (P. 27-2)

Enter data yourselves

Return again to Stata run-through

merge command

- Used when you want to add data to a pre-existing data set, or you have more than one dataset that has all the variables you need for analysis.
- Most important thing: each dataset must have (at least) one identifier that links observations, and allows merging.
- Second thing: both datasets must be sorted on the common identifier(s)

Example: one-for-one match

Election results, election_results.dta

county	cand1	cand2	cand2
A	10	20	30
B	40	50	60
C	70	80	90
Z	500	40	30

Demographics, demographics.dta

county	income	educ	catholic
A	10,000	.2	.3
B	40,000	.5	.6
C	70,000	.8	.9
Z	5,000	.95	.3

merge command results

- [assume both datasets have previously been sorted on county, by typing the command `sort county`]
- `use election_results.dta`
- `merge county using demographics.dta OR`
- `merge 1:1 county using demographics.dta`

Voila!

county	cand1	cand2	cand2	income	educ	catholic
A	10	20	30	10,000	.2	.3
B	40	50	60	40,000	.5	.6
C	70	80	90	70,000	.8	.9
Z	500	40	30	5,000	.95	.3

many-to-one merge

Demographic data, demographic_data.dta

county_code	town	income	education
A	Aville	50000	.3
A	Bobville	60000	.4
B	Candiceville	70000	.5
B	Dogville	80000	.5
C	Catville	100000	.5

County code mapping, county_code_mapping.dta

county_code	county_name
A	Adams
B	Brooks
C	Calhoun

merge command

- [make same sorting assumptions as before]
- `use demographic_data.dta`
- `merge m:1 county_code using county_code_mapping.dta`

Voila!

county_code	town	income	education	county_name
A	Aville	50000	.3	Adams
A	Bobville	60000	.4	Adams
B	Candiceville	70000	.5	Brooks
B	Dogville	80000	.5	Brooks
C	Catville	100000	.5	Calhoun

collapse command

county	DistrictName-en	voters	Paul	Bachmann	Johnson	Gingrich	Santorum	Huntsman	Other	Roemer	Romney	Perry	Cain
Adair	Adair - 1NW ADAIR	46	7	4		11	10	0	0	0	8	6	0
Adair	Adair - 2NE STUART	51	8	5		3	15	1	0	0	6	13	0
Adair	Adair - 3SW FONTANELLE	55	9	6		16	14	0	0	0	3	7	0
Adair	Adair - 4SE ORIENT	50	4	6		6	15	0	0	0	13	6	0
Adair	Adair - 5GF GREENFIELD	67	14	5		8	12	0	0	0	13	15	0
Adams	Adams – Carbon	28	7	0		5	12	0	0	0	3	1	0
Adams	Adams - Corning 1A	19	7	0		1	6	0	0	0	4	1	0
Adams	Adams - Corning 1B	3	3	0		0	0	0	0	0	0	0	0
Adams	Adams - Corning 2A	9	2	0		2	0	0	0	0	5	0	0
Adams	Adams - Corning 2B	8	5	1		0	1	0	0	0	0	1	0
Adams	Adams - Corning 3A	12	4	0		0	0	0	0	0	6	2	0
Adams	Adams - Corning 3B	19	9	0		1	6	0	0	0	1	2	0
Adams	Adams – Nodaway	10	1	1		5	0	0	0	0	2	1	0
Adams	Adams – Prescott	32	21	3		2	1	0	0	0	3	2	0
Adams	Adams – Quincy	22	7	0		2	8	0	0	0	3	2	0
Adams	Adams - SE Adams	38	8	4		6	13	0	0	0	5	2	0
Allamakee	Allamakee - FV/TL/HF CITY	28	7	0		6	9	0	0	0	6	0	0
Allamakee	Allamakee - LF/CN/LS/LS CITY	64	20	2		21	7	0	0	0	4	10	0
Allamakee	Allamakee - PC/LT/WV CITY	42	20	0		7	9	0	0	0	5	1	0
Allamakee	Allamakee - PO/FK	20	4	1		5	3	0	0	0	6	1	0
Allamakee	Allamakee - PV CITY	35	7	1		2	3	0	0	0	21	1	0
Allamakee	Allamakee - UC/IA/NA CITY	31	4	1		3	4	0	0	0	16	3	0
Allamakee	Allamakee - UP/MK/FC/JF/LL	122	53	2		18	14	0	0	0	28	7	0
Allamakee	Allamakee - WK 1 CITY	33	8	1		6	12	0	0	0	5	1	0

collapse (sum) voters-Cain,by(county)

county	voters	Paul	Bachmann	Johnson	Gingrich	Santorum	Huntsman	Other	Roemer	Romney	Perry	Cain
Adair	269	42	26	0	44	66	1	0	0	43	47	0
Adams	200	74	9	0	24	47	0	0	0	32	14	0
Allamakee	518	157	18	0	82	77	0	0	0	155	28	0
Appanoose	537	77	25	0	71	174	1	0	12	87	90	0
Audubon	223	41	17	0	32	54	0	0	0	48	31	0
Benton	1042	202	66	0	121	290	5	1	0	184	168	4
Black Hawk	3642	870	262	0	596	783	29	0	4	835	259	1
Boone	1344	276	104	0	160	400	4	0	0	230	170	0
Bremer	933	194	57	0	98	215	14	2	0	246	105	0
Buchanan	459	66	40	0	77	133	1	2	0	78	62	0
Buena Vista	716	169	26	0	128	154	3	0	0	124	110	2
Butler	552	99	41	0	71	157	4	0	0	92	87	0
Calhoun	435	75	31	0	54	131	2	2	0	69	71	0
Carroll	716	133	32	0	145	168	2	0	1	146	85	1
Cass	674	116	32	0	147	170	2	0	0	141	66	0
Cedar	711	188	34	0	84	167	4	1	0	165	67	0
Cerro Gordo	1571	304	100	0	235	345	5	1	0	408	170	2
Cherokee	537	95	20	0	78	155	0	0	0	126	63	0
Chickasaw	443	142	14	0	53	72	3	0	0	85	74	0
Clarke	367	98	42	0	46	51	1	2	0	65	62	0
Clay	733	150	40	0	137	165	4	2	0	149	75	0
Clayton	625	205	28	0	72	122	1	0	0	116	81	0
Clinton	1384	295	62	0	149	354	9	0	0	437	73	5
Crawford	437	72	22	0	84	101	0	0	0	93	64	0

Do-files

- Do-files are the Stata scripting language to automate analysis.
- Here is how the first five lines of the Iowa exercise would look in a do-file:

```
#delimit;
insheet using iowa_example_csv.dat;
list;
generate paulpct08=paul08/tvotes08;
generate paulpct12=paul12/tvotes12;
```

MIT OpenCourseWare
http://ocw.mit.edu

17.871 Political Science Laboratory

Spring 2012

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.