Data Preparation & Descriptive Statistics

(ver. 2.7)

Oscar Torres-Reyna
Data Consultant
otorres@princeton.edu

Basic definitions…

For statistical analysis we think of ***data*** as a collection of different pieces of information or facts. These pieces of information are called variables. A ***variable*** is an identifiable piece of data containing one or more values. Those values can take the form of a number or text (which could be converted into number)

In the table below variables var1 thru var5 are a collection of seven values, 'id' is the identifier for each observation. This dataset has information for seven cases (in this case people, but could also be states, countries, etc) grouped into five variables.

id	var1	var2	var3	var4	var5
1	7.3	32.27	0.1	Yes	Male
2	8.28	40.68	0.56	No	Female
3	3.35	5.62	0.55	Yes	Female
4	4.08	62.8	0.83	Yes	Male
5	9.09	22.76	0.26	No	Female
6	8.15	90.85	0.23	Yes	Female
7	7.59	54.94	0.42	Yes	Male

Data structure…

For data analysis your data should have variables as columns and observations as rows. The first row should have the column headings. Make sure your dataset has ***at least*** one identifier (for example, individual id, family id, etc.)

id	var1	var2	var3	var4	var5
1	7.3	32.27	0.1	Yes	Male
2	8.28	40.68	0.56	No	Female
3	3.35	5.62	0.55	Yes	Female
4	4.08	62.8	0.83	Yes	Male
5	9.09	22.76	0.26	No	Female
6	8.15	90.85	0.23	Yes	Female
7	7.59	54.94	0.42	Yes	Male

First row should have the variable names

Cross-sectional data

At least one identifier

Cross-sectional time series data or panel data

	id	year	var1	var2	var3
Group 1	1	2000	7	74.03	0.55
	1	2001	2	4.6	0.44
	1	2002	2	25.56	0.77
Group 2	2	2000	7	59.52	0.05
	2	2001	2	16.95	0.94
	2	2002	9	1.2	0.08
Group 3	3	2000	9	85.85	0.5
	3	2001	3	98.85	0.32
	3	2002	3	69.2	0.76

NOTE: See: http://www.statistics.com/resources/glossary/c/crossdat.php

Data format (ASCII)…

ASCII (American Standard Code for Information Interchange). The most universally accepted format. Practically any statistical software can open/read these type of files. Available formats:

- Delimited. Data is separated by comma, tab or space. The most common extension is *.csv (comma-separated value). Another type of extensions are *.txt for tab-separated data and *.prn for space-separated data. Any statistical package can read these formats.

- Record form (or fixed). Data is structured by fixed blocks (for example, var1 in columns 1 to 5, var2 in column 6 to 8, etc). You will need a codebook and to write a program (either in Stata, SPSS or SAS) to read the data. Extensions for the datasets could be *.dat, *.txt. For data in this format no column headings is available.

Data formats (comma-separated)…

Comma-separated value (*.csv)

```
ID,Last Name,First Name,City,State,Gender,Student Status,Major,Country,Age,SAT,Average score (grade),Height (in),Newspaper readership (times/wk),,,,,,,,,
1,DOE01,JANE01,Los Angeles,California,Female,Graduate,Politics,US,30,2263,67,61,5,,,,,,,,,
2,DOE02,JANE02,Sedona,Arizona,Female,Undergraduate,Math,US,19,2006,63,64,7,,,,,,,,,
3,DOE01,JOE01,Elmira,New York,Male,Graduate,Math,US,26,2221,78,73,6,,,,,,,,,
4,DOE02,JOE02,Lackawana,New York,Male,Graduate,Econ,US,33,1716,78,68,3,,,,,,,,,
5,DOE03,JOE03,Defiance,Ohio,Male,Graduate,Econ,US,37,1701,65,71,6,,,,,,,,,
6,DOE04,JOE04,Tel Aviv,Israel,Male,Graduate,Econ,Israel,25,1786,69,67,5,,,,,,,,,
7,DOE05,JOE05,Cimax,North Carolina,Male,Graduate,Politics,US,39,1577,96,70,5,,,,,,,,,
8,DOE03,JANE03,Liberal,Kansas,Female,Undergraduate,Politics,US,21,1842,87,62,5,,,,,,,,,
9,DOE04,JANE04,Montreal,Canada,Female,Undergraduate,Math,Canada,18,1813,91,62,6,,,,,,,,,
10,DOE05,JANE05,New York,New York,Female,Graduate,Math,US,33,2041,71,66,5,,,,,,,,,
11,DOE06,JOE06,Hot Coffe,Mississippi,Male,Undergraduate,Econ,US,18,1787,82,67,3,,,,,,,,,
12,DOE06,JANE06,Java,Virginia,Female,Graduate,Math,US,38,1513,79,59,5,,,,,,,,,
13,DOE07,JOE07,Varna,Bulgaria,Male,Graduate,Politics,Bulgaria,30,1637,79,63,4,,,,,,,,,
14,DOE08,JOE08,Moscow,Russia,Male,Graduate,Politics,Russia,30,1512,70,75,6,,,,,,,,,
15,DOE07,JANE07,Drunkard Creek,New York,Female,Undergraduate,Math,US,21,1338,82,64,5,,,,,,,,,
16,DOE08,JANE08,Mexican Hat,Utah,Female,Undergraduate,Econ,US,18,1821,80,63,3,,,,,,,
17,DOE09,JANE09,Amsterdam,Holland,Female,Undergraduate,Math,Holland,19,1494,75,60,3,,,,,,,
18,DOE10,JANE10,Mexico,Mexico,Female,Graduate,Politics,Mexico,31,2248,95,59,4,,,,,,,
19,DOE11,JANE11,Caracas,Venezuela,Female,Undergraduate,Math,Venezuela,18,2252,92,68,5,,,,,,,
20,DOE09,JOE09,San Juan,Puerto Rico,Male,Graduate,Politics,US,33,1923,95,63,7,,,,,,,
21,DOE12,JANE12,Remote,Oregon,Female,Undergraduate,Econ,US,19,1727,67,62,7,,,,,,,
22,DOE10,JOE10,New York,New York,Male,Undergraduate,Econ,US,21,1872,82,73,4,,,,,,,
23,DOE13,JANE13,The X,Massachusetts ,Female,Graduate,Politics,US,25,1767,89,68,6,,,,,,,
24,DOE14,JANE14,Beijing,China,Female,Undergraduate,Math,China,18,1643,79,65,6,,,,,,,
25,DOE11,JOE11,Stockholm,Sweden,Male,Undergraduate,Politics,Sweden,19,1919,88,64,4,,,,,,,
26,DOE12,JOE12,Embarrass,Minnesota,Male,Graduate,Econ,US,28,1434,96,71,4,,,,,,,
27,DOE13,JOE13,Intercourse,Pennsylvania,Male,Undergraduate,Math,US,20,2119,88,71,5,,,,,,,
28,DOE15,JANE15,Loco,Oklahoma,Female,Undergraduate,Econ,US,20,2309,64,68,6,,,,,,,
29,DOE14,JOE14,Buenos Aires,Argentina,Male,Graduate,Politics,Argentina,30,2279,85,72,3,,,,,,,
30,DOE15,JOE15,Acme,Louisiana,Male,Undergraduate,Econ,US,19,1907,79,74,3,,,,,,,
```

Data format (tab/space separated)…

Tab - separated value (*.txt)

ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)	Height (in)	Newspaper readership (times/wk)
1	00E01	JANE01	Los Angeles	California	Female	Graduate	Politics	US	30	2263	67	61	5
2	00E02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63	64	7
3	00E01	JOE01	Elmira	New York	Male	Graduate	Math	US	26	2221	78	73	6
4	00E02	JOE02	Lackawana	New York	Male	Graduate	Econ	US	33	1716	78	68	3
5	00E03	JOE03	Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65	71	6
6	00E04	JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69	67	5
7	00E05	JOE05	Cimax	North Carolina	Male	Graduate	Politics	US	39	1577	96	70	5
8	00E03	JANE03	Liberal	Kansas	Female	Undergraduate	Politics	US	21	1842	87	62	5
9	00E04	JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada	18	1813	91	62	6
10	00E05	JANE05	New York	New York	Female	Graduate	Math	US	33	2041	71	66	5
11	00E06	JOE06	Hot Coffe	Mississippi	Male	Undergraduate	Econ	US	18	1787	82	67	3
12	00E06	JANE06	Java	Virginia	Female	Graduate	Math	US	38	1513	79	59	5
13	00E07	JOE07	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria		30	1637	79	63	4
14	00E08	JOE08	Moscow	Russia	Male	Graduate	Politics	Russia	30	1512	70	75	6
15	00E07	JANE07	Drunkard Creek	New York	Female	Undergraduate	Math	US	21	1338	82	64	5
16	00E08	JANE08	Mexican Hat	Utah	Female	Undergraduate	Econ	US	18	1821	80	63	3
17	00E09	JANE09	Amsterdam	Holland	Female	Undergraduate	Math	Holland	19	1494	75	60	3
18	00E10	JANE10	Mexico	Mexico	Female	Graduate	Politics	Mexico	31	2248	95	59	4
19	00E11	JANE11	Caracas	Venezuela	Female	Undergraduate	Math	Venezuela		18	2252	92	68	5
20	00E09	JOE09	San Juan	Puerto Rico	Male	Graduate	Politics	US	33	1923	95	63	7
21	00E12	JANE12	Remote	Oregon	Female	Undergraduate	Econ	US	19	1727	67	62	7
22	00E10	JOE10	New York	New York	Male	Undergraduate	Econ	US	21	1872	82	73	4
23	00E13	JANE13	The X	Massachusetts	Female	Graduate	Politics	US	25	1767	89	68	6
24	00E14	JANE14	Beijing	China	Female	Undergraduate	Math	China	18	1643	79	65	6
25	00E11	JOE11	Stockholm	Sweden	Male	Undergraduate	Politics	Sweden	19	1919	88	64	4
26	00E12	JOE12	Embarrass	Minnesota	Male	Graduate	Econ	US	28	1434	96	71	4
27	00E13	JOE13	Intercourse	Pennsylvania	Male	Undergraduate	Math	US	20	2119	88	71	5
28	00E15	JANE15	Loco	Oklahoma	Female	Undergraduate	Econ	US	20	2309	64	68	6
29	00E14	JOE14	Buenos Aires	Argentina	Male	Graduate	Politics	Argentina		30	2279	85	72	3
30	00E15	JOE15	Acme	Louisiana	Male	Undergraduate	Econ	US	19	1907	79	74	3

Space - separated value (*.prn)

ID	Last Name	First Name	City	State	Gender	Student Statu	Major	Country	Age	SAT	Average score	Height (in)	Newspaper readership (times
100E01		JANE01	Los Angeles	California	Female	Graduate	Politics	US	30	2263	67	61	5
200E02		JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63	64	7
300E01		JOE01	Elmira	New York	Male	Graduate	Math	US	26	2221	78	73	6
400E02		JOE02	Lackawana	New York	Male	Graduate	Econ	US	33	1716	78	68	3
500E03		JOE03	Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65	71	6
600E04		JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69	67	5
700E05		JOE05	Cimax	North Carolin	Male	Graduate	Politics	US	39	1577	96	70	5
800E03		JANE03	Liberal	Kansas	Female	Undergraduate	Politics	US	21	1842	87	62	5
900E04		JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada	18	1813	91	62	6
1000E05		JANE05	New York	New York	Female	Graduate	Math	US	33	2041	71	66	5
1100E06		JOE06	Hot Coffe	Mississippi	Male	Undergraduate	Econ	US	18	1787	82	67	3
1200E06		JANE06	Java	Virginia	Female	Graduate	Math	US	38	1513	79	59	5
1300E07		JOE07	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	30	1637	79	63	4
1400E08		JOE08	Moscow	Russia	Male	Graduate	Politics	Russia	30	1512	70	75	6
1500E07		JANE07	Drunkard Cree	New York	Female	Undergraduate	Math	US	21	1338	82	64	5
1600E08		JANE08	Mexican Hat	Utah	Female	Undergraduate	Econ	US	18	1821	80	63	3
1700E09		JANE09	Amsterdam	Holland	Female	Undergraduate	Math	Holland	19	1494	75	60	3
1800E10		JANE10	Mexico	Mexico	Female	Graduate	Politics	Mexico	31	2248	95	59	4
1900E11		JANE11	Caracas	Venezuela	Female	Undergraduate	Math	Venezuela	18	2252	92	68	5
2000E09		JOE09	San Juan	Puerto Rico	Male	Graduate	Politics	US	33	1923	95	63	7
2100E12		JANE12	Remote	Oregon	Female	Undergraduate	Econ	US	19	1727	67	62	7
2200E10		JOE10	New York	New York	Male	Undergraduate	Econ	US	21	1872	82	73	4
2300E13		JANE13	The X	Massachusetts	Female	Graduate	Politics	US	25	1767	89	68	6
2400E14		JANE14	Beijing	China	Female	Undergraduate	Math	China	18	1643	79	65	6
2500E11		JOE11	Stockholm	Sweden	Male	Undergraduate	Politics	Sweden	19	1919	88	64	4
2600E12		JOE12	Embarrass	Minnesota	Male	Graduate	Econ	US	28	1434	96	71	4
2700E13		JOE13	Intercourse	Pennsylvania	Male	Undergraduate	Math	US	20	2119	88	71	5
2800E15		JANE15	Loco	Oklahoma	Female	Undergraduate	Econ	US	20	2309	64	68	6
2900E14		JOE14	Buenos Aires	Argentina	Male	Graduate	Politics	Argentina	30	2279	85	72	3
3000E15		JOE15	Acme	Louisiana	Male	Undergraduate	Econ	US	19	1907	79	74	3

Data format (record/fixed)…

Record form (fixed) ASCII (*.txt, *.dat). For this format you need a ***codebook*** to figure out the layout of the data (it indicates where a variable starts and where it ends). See next slide for an example. Notice that <u>fixed datasets do not have column headings</u>.

```
DOE01JANE01Los AngelesCaliforniaFemaleGraduatePoliticsUS302263676152DOE02JANE02SedonaArizonaFemaleUndergraduateMathUS192006636473
DOE01JOE01ElmiraNew YorkMaleGraduateMathUS262221787364DOE02JOE02LackawanaNew YorkMaleGraduateEconUS3317716786835
DOE03JOE03DefianceOhioMaleGraduateEconUS37170165716 6DOE04JOE04Tel AvivIsraelMaleGraduateEconIsrael251786696757
DOE05JOE05CimaxNorth CarolinaMaleGraduatePoliticsUS391577967058
DOE03JANE03LiberalKansasFemaleUndergraduatePoliticsUS211842876259DOE04JANE04MontrealCanadaFemaleUndergraduateMathCanada1818139162
10
DOE05JANE05New YorkNew YorkFemaleGraduateMathUS3320417166511DOE06JOE06Hot CoffeMississippiMaleUndergraduateEconUS1817878267312
DOE06JANE06JavaVirginiaFemaleGraduateMathUS3815137959513
DOE07JOE07VarnaBulgariaMaleGraduatePoliticsBulgaria30163779634 4DOE08JOE08MoscowRussiaMaleGraduatePoliticsRussia3015127075615
DOE07JANE07Drunkard CreekNew YorkFemaleUndergraduateMathUS211338826 4516DOE08JANE08Mexican HatUtahFemaleUndergraduateEconUS181821
8063317
DOE09JANE09AmsterdamHollandFemaleUndergraduateMathHolland1914947560318DOE10JANE10MexicoMexicoFemaleGraduatePoliticsMexico31224895
9419
DOE11JANE11CaracasVenezuelaFemaleUndergraduateMathVenezuela1822529268520
DOE09JOE09San JuanPuerto RicoMaleGraduatePoliticsUS3319239563721DOE12JANE12RemoteOregonFemaleUndergraduateEconUS1917276762722
DOE10JOE10New YorkNew YorkMaleUndergraduateEconUS2118728273423
DOE13JANE13The XMassachusetts FemaleGraduatePoliticsUS2517678968624DOE14JANE14BeijingChinaFemaleUndergraduateMathChina18164379656
5
DOE11JOE11StockholmSwedenMaleUndergraduatePoliticsSweden191919886 4426DOE12JOE12EmbarrassMinnesotaMaleGraduateEconUS2814349671427
DOE13JOE13IntercoursePennsylvaniaMaleUndergraduateMathUS2021198871528DOE15JANE15LocoOklahomaFemaleUndergraduateEconUS20230964686
29DOE14JOE14Buenos AiresArgentinaMaleGraduatePoliticsArgentina3022798572330DOE15JOE15AcmeLouisianaMaleUndergraduateEconUS19190779
43
```

Codebook (ASCII to Stata using `infix`)

NOTE: The following is a small example of a codebook. Codebooks are like maps to help you figure out the structure of the data. Codebooks differ on how they present the layout of the data, in general, you need to look for: variable name, start column, end column or length, and format of the variable (whether is numeric and how many decimals (identified with letter 'F') or whether is a string variable marked with letter 'A')

Data Locations

Variable	Rec	Start	End	Format
var1	1	1	7	F7.2
var2	1	24	25	F2.0
var3	1	26	27	A2
var4	1	32	33	F2.0
var5	1	44	45	A2

In Stata you write the following to open the dataset. In the command window type:

```
infix var1 1-7 var2 24-25 str2 var3 26-27 var4 32-33 str2 var5 44-45 using mydata.dat
```

Notice the '`str#`' before var3 and var5, this is to indicate that these variables are string (text). The number in `str` refers to the length of the variable.

If you get an error like …`cannot be read as a number for`… click here

From ASCII to Stata using a dictionary file/`infile`

Using notepad or the do-file editor type:

```
dictionary using c:\data\mydata.dat {
        _column(1)        var1      %7.2f     "Label for var1"
        _column(24)       var2      %2f       "Label for var2"
        _column(26) str2  var3      %2s       "Label for var3"
        _column(32)       var4      %2f       "Label for var4"
        _column(44) str2  var5      %2s       "Label for var5"
    }
/*Do not forget to close the brackets and press enter after the last bracket*/
```

Notice that the numbers in `_column(#)` refers to the position where the variable starts based on what the codebook shows. The option '`str#`' indicates that the variable is a string (text or alphanumeric) with two characters, here you need to specify the length of the variable for Stata to read it correctly.

Save it as *`mydata.dct`*

To read data using the dictionary we need to import the data by using the command `infile`. If you want to use the menu go to File – Import - "ASCII data in fixed format with a data dictionary".

With `infile` we run the dictionary by typing:

```
infile using c:\data\mydata
```

NOTE: Stata commands sometimes do not work with copy-and-paste. If you get error try re-typing the commands

If you get an error like …`cannot be read as a number for…` click here

From ASCII to Stata using a dictionary file/`infile` (data with more than one record)

If your data is in more than one records using notepad or the do-file editor type:

```
dictionary using c:\data\mydata.dat {
       _lines(2)
       _line(1)
       _column(1)       var1      %7.2f     "Label for var1"
       _column(24)      var2      %2f       "Label for var2"
       _line(2)
       _column(26) str2 var3      %2s       "Label for var3"
       _column(32)      var4      %2f       "Label for var4"
       _column(44) str2 var5      %2s       "Label for var5"
 }
/*Do not forget to close the brackets and press enter after the last bracket*/
```

Notice that the numbers in `_column(#)` refers to the position where the variable starts based on what the codebook shows.

Save it as *mydata.dct*

To read data using the dictionary we need to import the data by using the command `infile`. If you want to use the menu go to File – Import - "ASCII data in fixed format with a data dictionary".

With `infile` we run the dictionary by typing:

```
infile using c:\data\mydata
```

NOTE: Stata commands sometimes do not work with copy-and-paste. If you get error try re-typing the commands
For more info on data with records see http://www.columbia.edu/cu/lweb/indiv/dssc/eds/stata_write.html

If you get an error like …`cannot be read as a number for`… click here

From ASCII to Stata: error message

If running `infix` or `infile` you get errors like:

```
'1-1001-' cannot be read as a number for var1[14]
'de111' cannot be read as a number for var2[11]
'xvet-' cannot be read as a number for var3[15]
'0---0' cannot be read as a number for var4[16]
'A5' cannot be read as a number for var5[16]
```

Make sure you specified those variables to be read as strings (`str`) and set to the correct length (`str#`), see the codebook for these.

Double-check the data locations from the codebook. If the data file has more than one record make sure is indicated in the dictionary file.

If after checking for the codebook you find no error in the data locations or the data type, then depending of the type of variable, this may or may not be an error. Stata will still read the variables but those non-numeric observations will be set to missing.

From ASCII to SPSS

Using the syntax editor in SPSS and following the data layout described in the codebook, type:

```
FILE HANDLE FHAND /NAME='C:\data\mydata.dat' /LRECL=1003.
DATA LIST FILE=FHAND FIXED RECORDS = 1 TABLE /
   var1 1-7
   var2 24-25
   var3 26-27 (A)
   var4 32-33
   var5 44-45 (A).
EXECUTE.
```

You get /`LRECL` from the codebook.

Select the program and run it by clicking on the arrow 

If you have more than one record type:

```
FILE HANDLE FHAND /NAME='C:\data\mydata.dat' /LRECL=1003.
DATA LIST FILE=FHAND FIXED RECORDS = 2 TABLE
/1
   var1 1-7
   var2 24-25
   var3 26-27 (A)
/2
   var4 32-33
   var5 44-45 (A).
EXECUTE.
```


Notice the `'(A)'` after var3 and var5, this is to indicate that these variables are string (text).

From SPSS/SAS to Stata

If your data is already in SPSS format (*.sav) or SAS(*.sas7bcat).You can use the command `usespss` to read SPSS files in Stata or the command `usesas` to read SAS files.

If you have a file in SAS XPORT format you can use `fduse` (or go to file-import).

For SPSS and SAS, you may need to install it by typing

```
ssc install usespss
ssc install usesas
```

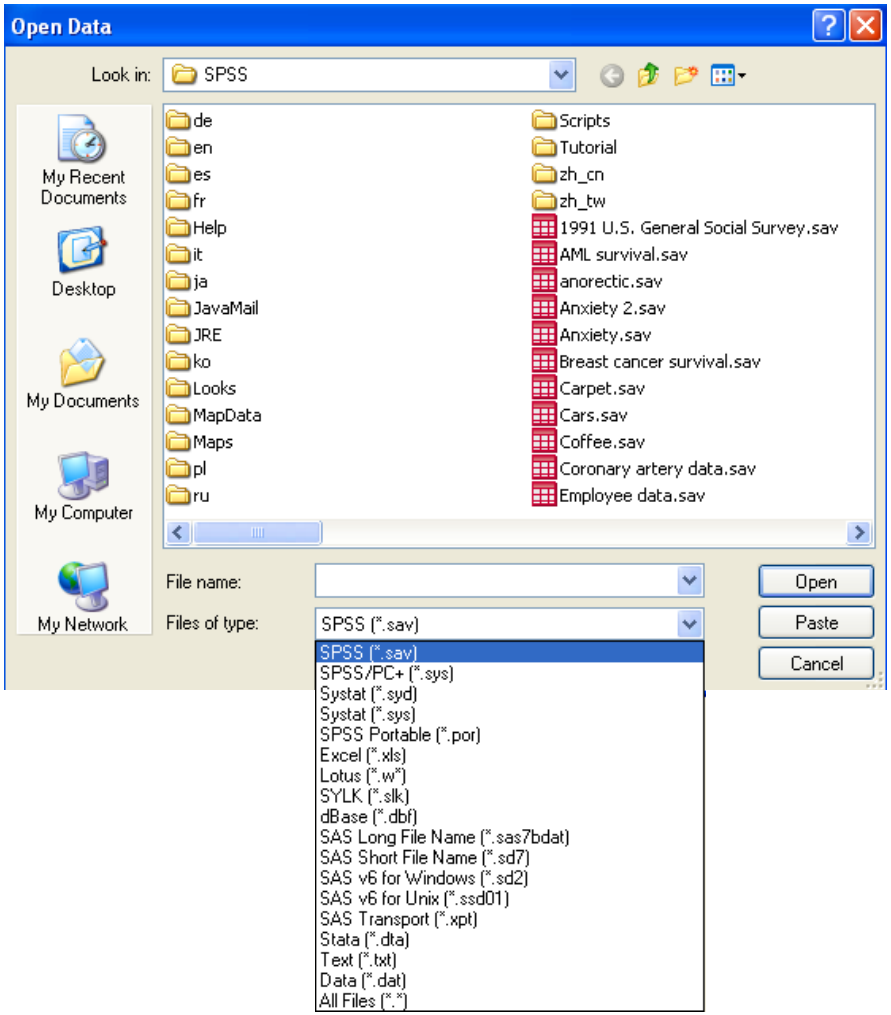
Once installed just type

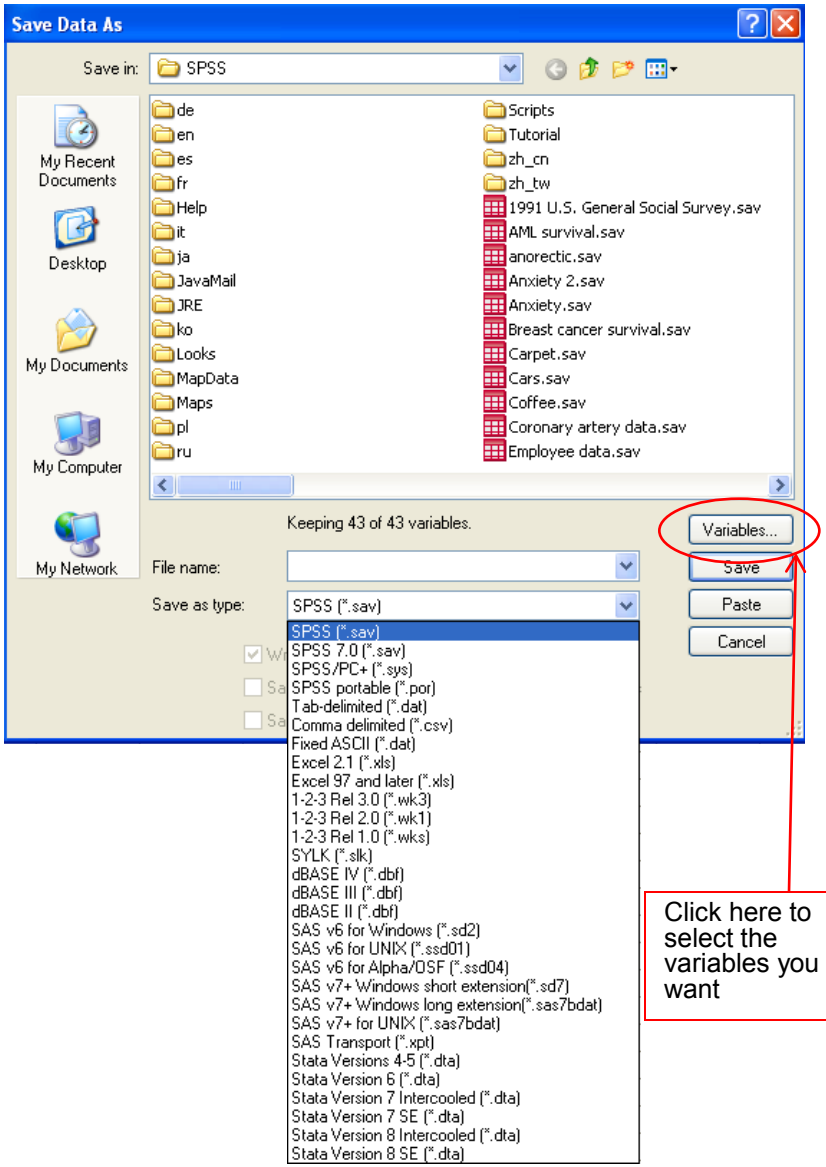
```
usespss using "c:\mydata.sav"
usesas using "c:\mydata.sas7bcat"
```

Type `help usespss` or `help usesas` for more details.

Loading data in SPSS

SPSS can read/save-as many proprietary data formats, go to file-open-data or file-save as





Loading data in R

1. tab-delimited (*.txt), type:

```
mydata <- read.table("mydata.txt")
mydata <- read.table("mydata.txt", header = TRUE, na.strings = "-9") #If
    missing data is coded as "-9"
```

2. space-delimited (*.prn), type:

```
mydata <- read.table("mydata.prn")
```

3. comma-separated value (*.csv), type:

```
mydata <- read.csv("mydata.csv")
mydata <- read.csv("mydata.csv", header = TRUE) #With column headings
```

4. From SPSS/Stata to R use the `foreign package`, type:

```
library(foreign)  # Load the foreign package.
stata.data <- read.dta("mydata.dta") # For Stata.
spss.data <- read.spss("mydata.sav", to.data.frame = TRUE) # For SPSS.
```

5. To load data in R format use

```
mydata <- load("mydata.RData")
```

Source: http://gking.harvard.edu/zelig/docs/static/syntax.pdf

Also check: http://www.ats.ucla.edu/stat/R/modules/raw_data.htm

Other data formats…

Features	Stata	SPSS	SAS	R
Data extensions	***.dta**	***.sav, *.por (portable file)**	***.sas7bcat, *.sas#bcat, *.xpt (xport files)**	***.Rdata**
User interface	Programming/point-and-click	Mostly point-and-click	Programming	Programming
Data manipulation	Very strong	Moderate	Very strong	Very strong
Data analysis	Powerful	Powerful	Powerful/versatile	Powerful/versatile
Graphics	Very good	Very good	Good	Good
Cost	Affordable (perpetual licenses, renew only when upgrade)	Expensive (but not need to renew until upgrade, long term licenses)	Expensive (yearly renewal)	Open source
Program extensions	*.do (do-files)	*.sps (syntax files)	*.sas	*.txt (log files)
Output extension	*.log (text file, any word processor can read it), *.smcl (formated log, only Stata can read it).	*.spo (only SPSS can read it)	(various formats)	*.txt (log files, any word processor can read)

Compress data files (*.zip, *.gz)

If you have datafiles with extension *.zip, *.gz, *.rar you need file compression software to extract the datafiles. You can use Winzip, WinRAR or 7-zip among others.

7-zip (http://7-zip.org/) is freeware and deals with most compressed formats.

Stata allows you to unzip files from the command window.

```
unzipfile "c:\data\mydata.zip"
```

You can also zip file using `zipfile`

```
zipfile myzip.zip mydata.dta
```

Before you start

Once you have your data in the proper format, before you perform any analysis you need to explore and prepare it first:

1. Make sure variables are in columns and observations in rows.
2. Make sure you have all variables you need.
3. Make sure there is at least one id.
4. If times series make sure you have the years you want to include in your study.
5. Make sure missing data has either a blank space or a dot (‘.’)
6. Make sure to make a back-up copy of your original dataset.
7. Have the codebook handy.

Stata color-coded system

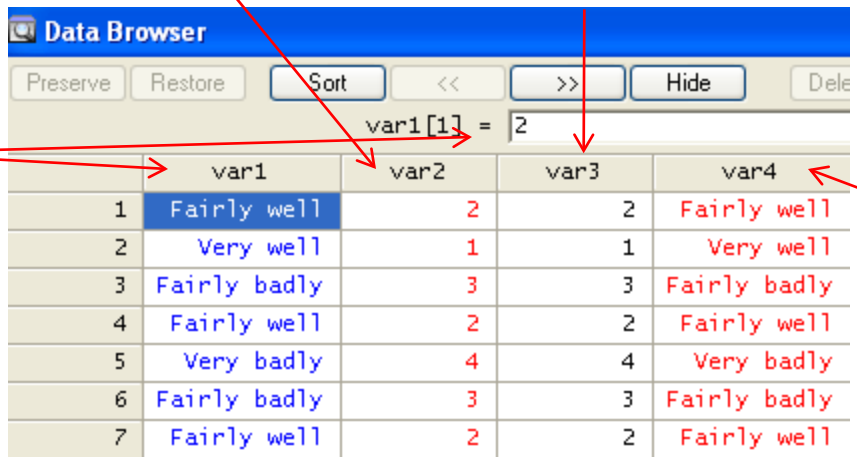
An important step is to make sure variables are in their expected format. Numeric should be numeric and text should be text.

Stata has a color-coded system for each type. Black is for numbers, red is for text or string and blue is for labeled variables.

Var2 is a string variable even though you see numbers. You can't do any statistical procedure with this variable other than simple frequencies

Var3 is a numeric You can do any statistical procedure with this variable

For var1 a value 2 has the label "Fairly well". It is still a numeric variable



	var1	var2	var3	var4
1	Fairly well	2	2	Fairly well
2	Very well	1	1	Very well
3	Fairly badly	3	3	Fairly badly
4	Fairly well	2	2	Fairly well
5	Very badly	4	4	Very badly
6	Fairly badly	3	3	Fairly badly
7	Fairly well	2	2	Fairly well

Var4 is clearly a string variable. You can do frequencies and crosstabulations with this but not any statistical procedure.

Cleaning your variables

If you are using datasets with categorical variables you need to clean them by getting rid of the non-response categories like ‘do not know’, ‘no answer’, ‘no applicable’, ‘not sure’, ‘refused’, etc.

Usually non-response categories have higher values like 99, 999, 9999, etc (or in some cases negative values). Leaving these will bias, for example, the mean age or your regression results as outliers.

In the example below the non-response is coded as 999 and if we leave this the mean age would be 80 years, removing the 999 and setting it to missing, the average age goes down to 54 years.

This is a frequency of age, notice the 999 value for the no response.

88	2	0.15	96.58
90	3	0.22	96.80
92	4	0.29	97.09
93	1	0.07	97.16
95	1	0.07	97.23
999	38	2.77	100.00
Total	1,373	100.00	

In Stata you can type

```
replace age=. if age==999
```

or

```
replace age=. if age>100
```

```
. tabstat age age_w999
```

stats	age	age_w999
mean	54.58801	80.72615

Cleaning your variables

No response categories not only affect the statistics of the variable, it may also affect the interpretation and coefficients of the variable if we do not remove them.

In the example below responses go from ‘very well’ to ‘refused’, with codes 1 to 6. Leaving the variable ‘as-is’ in a regression model will misinterpret the variable as going from quite positive to … refused? This does not make sense. You need to clean the variable by eliminating the no response so it goes from positive to negative. Even more, you may have to reverse the valence so the variable goes from negative to positive for a better/easier interpretation.

```
. tab var1
```

Status of Nat'l Eco	Freq.	Percent	Cum.
Very well	149	10.85	10.85
Fairly well	670	48.80	59.65
Fairly badly	348	25.35	85.00
Very badly	191	13.91	98.91
Not sure	12	0.87	99.78
Refused	3	0.22	100.00
Total	1,373	100.00	

=

```
. tab var1, nolabel
```

Status of Nat'l Eco	Freq.	Percent	Cum.
1	149	10.85	10.85
2	670	48.80	59.65
3	348	25.35	85.00
4	191	13.91	98.91
5	12	0.87	99.78
6	3	0.22	100.00
Total	1,373	100.00	

Cleaning your variables (using `recode` in Stata)

First, **never** work with the original variable, *always keep originals original*.

The command `recode` in Stata lets you create a new variable without modifying the original.

```
recode var1 (1=4 "Very well") (2=3 "Fairly well") (3=2 "Fairly badly")
(4=1 "Very badly") (else=.), gen(var1_rec) label(var1_rec)
```

Get frequencies of both variables: `var1` and `var1_rec` to verify:

```
. tab var1
```

Status of Nat'l Eco	Freq.	Percent	Cum.
Very well	149	10.85	10.85
Fairly well	670	48.80	59.65
Fairly badly	348	25.35	85.00
Very badly	191	13.91	98.91
Not sure	12	0.87	99.78
Refused	3	0.22	100.00
Total	1,373	100.00	

```
. tab  var1_rec
```

RECODE of var1 (Status of Nat'l Eco)	Freq.	Percent	Cum.
Very badly	191	14.06	14.06
Fairly badly	348	25.63	39.69
Fairly well	670	49.34	89.03
Very well	149	10.97	100.00
Total	1,358	100.00	

Now you can use `var1_rec` in a regression since it is an ordinal variable where higher values mean positive opinions. This process is useful when combining variables to create indexes.

For additional help on data management, analysis and presentation please check:
http://dss.princeton.edu/training/
http://dss.princeton.edu/

Reshape wide to long (if original data in Excel)

The following dataset is not ready for analysis, years are in columns and cases and variables are in rows (click here to get it). The ideal is for years and countries to be in rows and variables (var1 and var2) in columns. We should have four columns: Country, Year, var1and var2

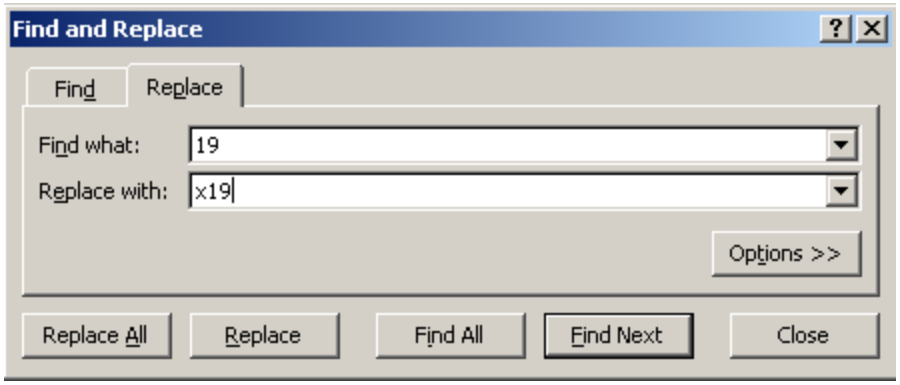
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Country**	**Variable**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**
2	A	var1	..	..	8000.01	8212.90	7847.36	7702.89	7288.48	6430.98	6932.45	7486.24	8094.17
3	A	var2	..	..	6.83	2.66	-4.45	-1.84	-5.38	-11.77	7.80	7.99	8.12
4	B	var1	18268.01	18738.99	19360.46	20151.42	20715.54	20866.90	21364.02	21801.41	22404.59	22676.26	23039.43
5	B	var2	2.87	2.58	3.32	4.09	2.80	0.73	2.38	2.05	2.77	1.21	1.60
6	C	var1	21088.14	21608.14	21988.64	22739.28	23436.61	24194.85	24300.57	24411.48	24650.02	25076.01	25346.01
7	C	var2	1.60	2.47	1.76	3.41	3.07	3.24	0.44	0.46	0.98	1.73	1.08
8	D	var1	313.74	321.36	331.76	342.12	351.70	365.33	377.15	386.26	398.86	415.96	432.63
9	D	var2	2.66	2.43	3.24	3.12	2.80	3.87	3.24	2.42	3.26	4.29	4.01
10	E	var1	21123.66	21659.55	22299.13	22972.31	23613.87	24150.86	24788.69	25368.87	25885.48	26582.19	26890.73
11	E	var2	2.69	2.54	2.95	3.02	2.79	2.27	2.64	2.34	2.04	2.69	1.16
12	F	var1	29941.64	30703.73	31716.04	32671.27	33748.21	34599.47	34483.98	34669.47	35312.75	36450.55	37267.33
13	F	var2	1.32	2.55	3.30	3.01	3.30	2.52	-0.33	0.54	1.86	3.22	2.24
14	G	var1	4891.60	5063.81	5328.88	5512.59	5647.06	5934.98	5864.12	5852.99	5872.29	6055.92	6162.84
15	G	var2	-7.86	3.52	5.23	3.45	2.44	5.10	-1.19	-0.19	0.33	3.13	1.77

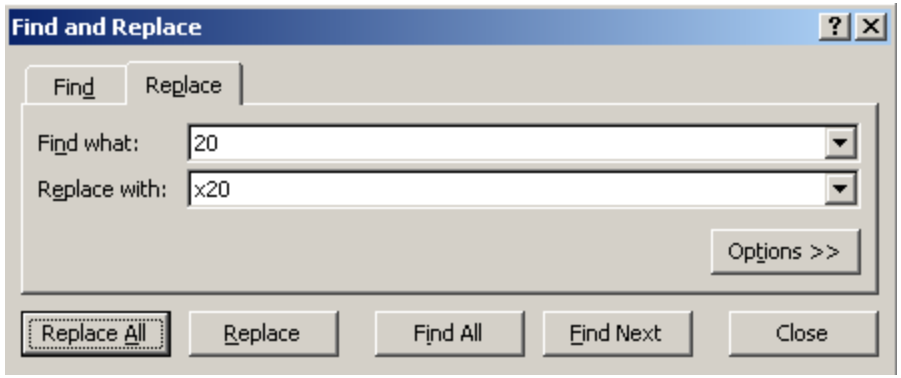
We can prepare this dataset using Stata but we need to do some changes in Excel.

For R please see: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/reshape.html

Reshape wide to long (if original data in Excel)

First, you need to add a character to the column headings so Stata can read them. Stata does not take numbers as variable names. In this case we add an “x” to the years. In excel you do this by using the ‘replace’ function. For the 1900s we replace “19” for “x19”, same for the 2000s (make sure to select only the headings). See the following



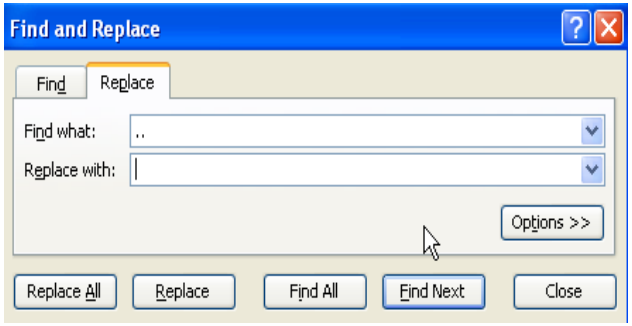


Reshape wide to long (if original data in Excel)

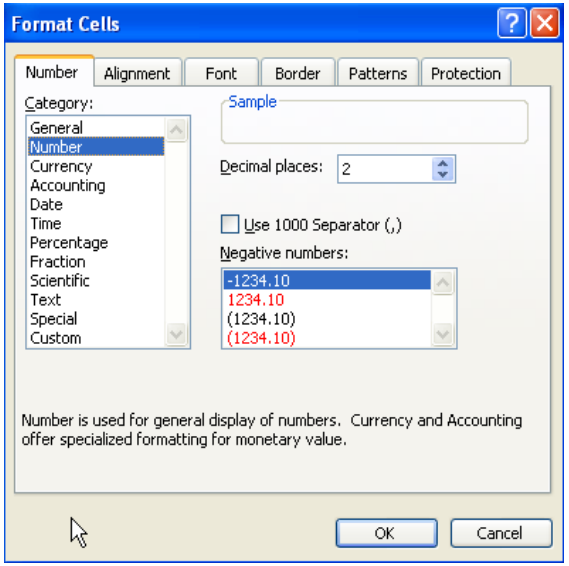
We have…

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Country**	**Variable**	**x1995**	**x1996**	**x1997**	**x1998**	**x1999**	**x2000**	**x2001**	**x2002**	**x2003**	**x2004**	**x2005**
2	A	var1	..	..	8000.01	8212.90	7847.36	7702.89	7288.48	6430.98	6932.45	7486.24	8094.17
3	A	var2	..	..	6.83	2.66	-4.45	-1.84	-5.38	-11.77	7.80	7.99	8.12
4	B	var1	18268.01	18738.99	19360.46	20151.42	20715.54	20866.90	21364.02	21801.41	22404.59	22676.26	23039.43
5	B	var2	2.87	2.58	3.32	4.09	2.80	0.73	2.38	2.05	2.77	1.21	1.60
6	C	var1	21088.14	21608.14	21988.64	22739.28	23436.61	24194.85	24300.57	24411.48	24650.02	25076.01	25346.01
7	C	var2	1.60	2.47	1.76	3.41	3.07	3.24	0.44	0.46	0.98	1.73	1.08
8	D	var1	313.74	321.36	331.76	342.12	351.70	365.33	377.15	386.26	398.86	415.96	432.63
9	D	var2	2.66	2.43	3.24	3.12	2.80	3.87	3.24	2.42	3.26	4.29	4.01
10	E	var1	21123.66	21659.55	22299.13	22972.31	23613.87	24150.86	24788.69	25368.87	25885.48	26582.19	26890.73
11	E	var2	2.69	2.54	2.95	3.02	2.79	2.27	2.64	2.34	2.04	2.69	1.16
12	F	var1	29941.64	30703.73	31716.04	32671.27	33748.21	34599.47	34483.98	34669.47	35312.75	36450.55	37267.33
13	F	var2	1.32	2.55	3.30	3.01	3.30	2.52	-0.33	0.54	1.86	3.22	2.24
14	G	var1	4891.60	5063.81	5328.88	5512.59	5647.06	5934.98	5864.12	5852.99	5872.29	6055.92	6162.84
15	G	var2	-7.86	3.52	5.23	3.45	2.44	5.10	-1.19	-0.19	0.33	3.13	1.77

Replace the dots ".." (or any string character) with a blank



Make sure the numbers are numbers. Select all and format cells as numbers.

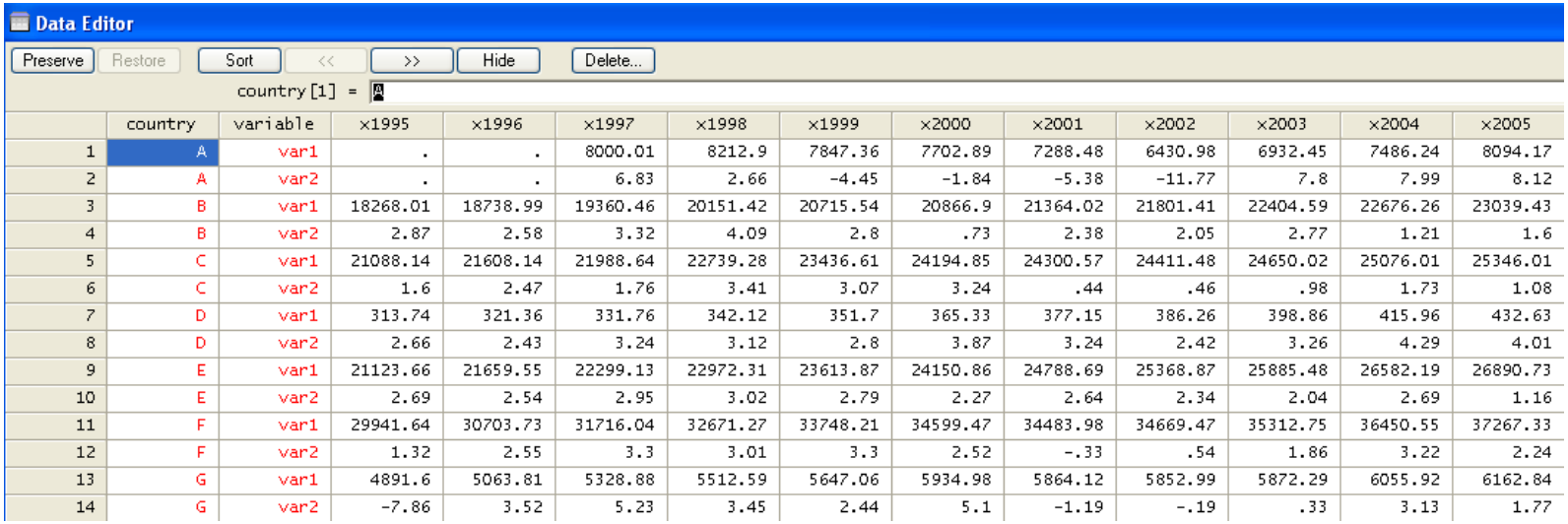


Reshape wide to long (from Excel to Stata)

The table should look like.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Country**	**Variable**	**x1995**	**x1996**	**x1997**	**x1998**	**x1999**	**x2000**	**x2001**	**x2002**	**x2003**	**x2004**	**x2005**
2	A	var1			8000.01	8212.90	7847.36	7702.89	7288.48	6430.98	6932.45	7486.24	8094.17
3	A	var2			6.83	2.66	-4.45	-1.84	-5.38	-11.77	7.80	7.99	8.12
4	B	var1	18268.01	18738.99	19360.46	20151.42	20715.54	20866.90	21364.02	21801.41	22404.59	22676.26	23039.43
5	B	var2	2.87	2.58	3.32	4.09	2.80	0.73	2.38	2.05	2.77	1.21	1.60
6	C	var1	21088.14	21608.14	21988.64	22739.28	23436.61	24194.85	24300.57	24411.48	24650.02	25076.01	25346.01
7	C	var2	1.60	2.47	1.76	3.41	3.07	3.24	0.44	0.46	0.98	1.73	1.08
8	D	var1	313.74	321.36	331.76	342.12	351.70	365.33	377.15	386.26	398.86	415.96	432.63
9	D	var2	2.66	2.43	3.24	3.12	2.80	3.87	3.24	2.42	3.26	4.29	4.01
10	E	var1	21123.66	21659.55	22299.13	22972.31	23613.87	24150.86	24788.69	25368.87	25885.48	26582.19	26890.73
11	E	var2	2.69	2.54	2.95	3.02	2.79	2.27	2.64	2.34	2.04	2.69	1.16
12	F	var1	29941.64	30703.73	31716.04	32671.27	33748.21	34599.47	34483.98	34669.47	35312.75	36450.55	37267.33
13	F	var2	1.32	2.55	3.30	3.01	3.30	2.52	-0.33	0.54	1.86	3.22	2.24
14	G	var1	4891.60	5063.81	5328.88	5512.59	5647.06	5934.98	5864.12	5852.99	5872.29	6055.92	6162.84
15	G	var2	-7.86	3.52	5.23	3.45	2.44	5.10	-1.19	-0.19	0.33	3.13	1.77

Copy and paste the table from Excel to Stata. In Stata go to Data -> Data Editor



	country	variable	x1995	x1996	x1997	x1998	x1999	x2000	x2001	x2002	x2003	x2004	x2005
1	A	var1	.	.	8000.01	8212.9	7847.36	7702.89	7288.48	6430.98	6932.45	7486.24	8094.17
2	A	var2	.	.	6.83	2.66	-4.45	-1.84	-5.38	-11.77	7.8	7.99	8.12
3	B	var1	18268.01	18738.99	19360.46	20151.42	20715.54	20866.9	21364.02	21801.41	22404.59	22676.26	23039.43
4	B	var2	2.87	2.58	3.32	4.09	2.8	.73	2.38	2.05	2.77	1.21	1.6
5	C	var1	21088.14	21608.14	21988.64	22739.28	23436.61	24194.85	24300.57	24411.48	24650.02	25076.01	25346.01
6	C	var2	1.6	2.47	1.76	3.41	3.07	3.24	.44	.46	.98	1.73	1.08
7	D	var1	313.74	321.36	331.76	342.12	351.7	365.33	377.15	386.26	398.86	415.96	432.63
8	D	var2	2.66	2.43	3.24	3.12	2.8	3.87	3.24	2.42	3.26	4.29	4.01
9	E	var1	21123.66	21659.55	22299.13	22972.31	23613.87	24150.86	24788.69	25368.87	25885.48	26582.19	26890.73
10	E	var2	2.69	2.54	2.95	3.02	2.79	2.27	2.64	2.34	2.04	2.69	1.16
11	F	var1	29941.64	30703.73	31716.04	32671.27	33748.21	34599.47	34483.98	34669.47	35312.75	36450.55	37267.33
12	F	var2	1.32	2.55	3.3	3.01	3.3	2.52	-.33	.54	1.86	3.22	2.24
13	G	var1	4891.6	5063.81	5328.88	5512.59	5647.06	5934.98	5864.12	5852.99	5872.29	6055.92	6162.84
14	G	var2	-7.86	3.52	5.23	3.45	2.44	5.1	-1.19	-.19	.33	3.13	1.77

NOTE: You can save the excel file as *.csv and open it in Stata typing `insheet using exceltable.csv`

Reshape wide to long (summary)

id	x2001	x2002	x2003
1	2	7	1
2	3	5	9
3	1	1	8

```
gen id = _n
order id
reshape long x , i(id) j(year)
```

id	year	x
1	1	2
1	2	7
1	3	1
2	1	3
2	2	5
2	3	9
3	1	1
3	2	1
3	3	8

date	x_var1	x_var2	x_var3
1	2	7	1
2	3	5	9
3	1	1	8

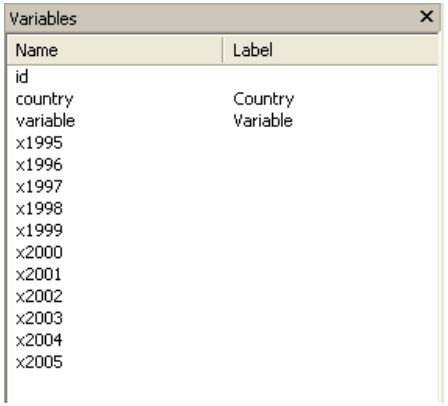
```
reshape long x_var , i(date) j(id) str
```

date	id	x_var
1	1	2
1	2	7
1	3	1
2	1	3
2	2	5
2	3	9
3	1	1
3	2	1
3	3	8

Reshape (Stata, 1)

Back to the example, create a unique id for each observation, type:

```
gen id = _n
order id
```



To reshape from wide to long, type

```
reshape long x, i(id) j(year)
```

```
. reshape long x, i(id) j(year)
(note: j = 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       14   ->     154
Number of variables                  14   ->       5
j variable (11 values)                    ->   year
xij variables:
                  x1995 x1996 ... x2005   ->   x
-----------------------------------------------------------------------------
```

Where:

- **long –** Goes from wide to long format.
- **x –** The variables with the prefix “x” (x1960, x1961, x1962, etc.) are to be converted from wide to long.
- **i(id) –** A unique identifier for the wide format is in variable “id”.
- **j(year) –** Indicates that the suffix of “x” (x1961, x1962, x1963, …), the years, should be put in variable called “year”.

NOTE: If you have more than one variable you can list them as follows:

```
reshape long x y z, i(id) j(year)
```

Reshape wide to long (Stata, 2)

The data it should look like the picture below. Notice that `var1` and `var2` are together in one column as variable `'x'` (the prefix we originally had for the years). If we had one variable we are done, in this example we have **two** and we need to separate them into two columns, `var1` and `var2` . Basically we need to reshape again but this time from long to wide.

	id	year	country	variable	x
1	1	1995	A	var1	.
2	1	1996	A	var1	.
3	1	1997	A	var1	8000.01
4	1	1998	A	var1	8212.9
5	1	1999	A	var1	7847.36
6	1	2000	A	var1	7702.89
7	1	2001	A	var1	7288.48
8	1	2002	A	var1	6430.98
9	1	2003	A	var1	6932.45
10	1	2004	A	var1	7486.24
11	1	2005	A	var1	8094.17
12	2	1995	A	var2	.
13	2	1996	A	var2	.
14	2	1997	A	var2	6.83
15	2	1998	A	var2	2.66
16	2	1999	A	var2	-4.45
17	2	2000	A	var2	-1.84
18	2	2001	A	var2	-5.38
19	2	2002	A	var2	-11.77
20	2	2003	A	var2	7.8
21	2	2004	A	var2	7.99

Reshape (Stata, 3)

To separate `var1` and `var2` we need to do a little bit of work.

First we need to create a new variable with the labels of each variable, type

```
encode variable, gen(varlabel)
```

```
. encode variable, gen(varlabel)

. tab  varlabel

   Variable |      Freq.     Percent        Cum.
------------+-----------------------------------
       var1 |         77       50.00       50.00
       var2 |         77       50.00      100.00
------------+-----------------------------------
      Total |        154      100.00

. tab  varlabel, nolabel

   Variable |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         77       50.00       50.00
          2 |         77       50.00      100.00
------------+-----------------------------------
      Total |        154      100.00
```

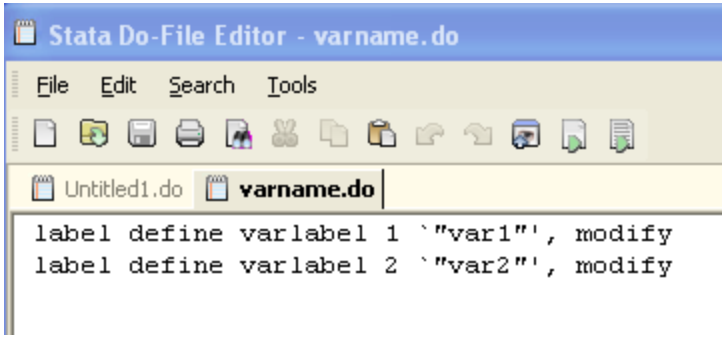
Create a do-file with the labels for each variable. This comes in handy when dealing with lots of variables.

```
label save varlabel using varname, replace
```

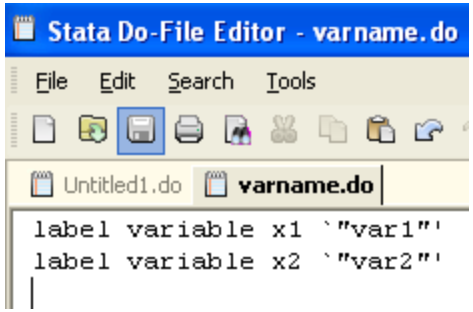
You will notice that a file `varname.do` is created.

```
. label save varlabel using varname, replace
file varname.do saved
```

Open the do-file with the do-file editor and do the following changes…



- Change “label define” to “label variable”
- Change “varlabel 1” to “x1” and “varlabel 2” to “x2”
- Delete “, modify
- Save the do-file



Reshape (Stata, 4)

To separate `var1` and `var2` we need to reshape again, this time from long to wide. First we need to create another id to identify the groups (country and years), type

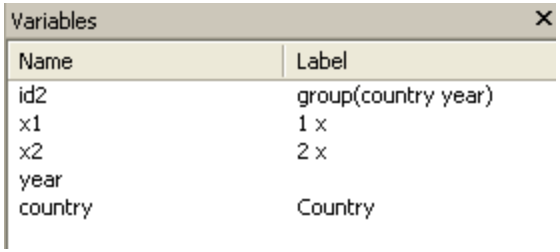
```
egen id2 = group(country year)
move id2 year
drop id
drop variable
```

Reshape the data by typing

```
reshape wide x, i(id2) j(varlabel)
order  id2 country year x1 x2
```

```
. reshape wide x, i(id2) j(varlabel)
(note: j = 1 2)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                      154   ->     77
Number of variables                   5   ->      5
j variable (2 values)          varlabel   ->   (dropped)
xij variables:
                                      x   ->   x1 x2
-----------------------------------------------------------------------------
```



Where:

wide – Indicates long to wide format.
x – The variable of interest to go from long to wide is called "data".
i(id2) – A unique identifier for the wide format is in variable "id2".
j(varlabel) – Indicates that the suffix of "data" has to be taken from ""varlabel" ("varlabel" has two categories: 1 –var1- and 2 – var2).

NOTE: If "j" is not available in your dataset, you may be able to generate one using the following command:

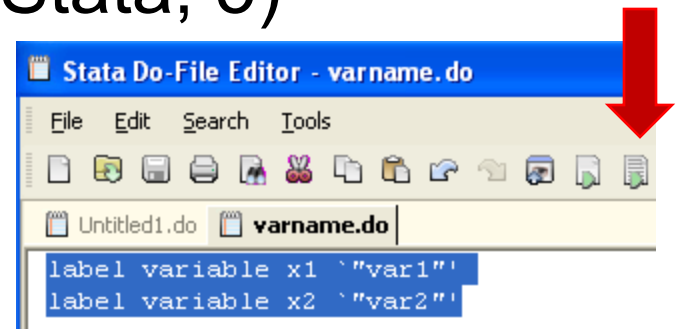
```
bysort id: gen jvar=_n
```

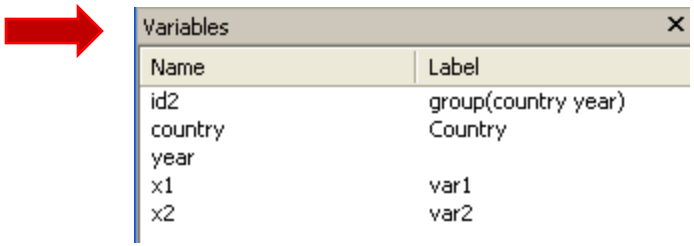
Then reshape

```
reshape wide data, i(id) j(jvar)
```

Reshape (Stata, 5)

Run the do-file `varname.do` by selecting all and clicking on the last icon, this will change the labels for x1 and x2





The final dataset will look like…

	id2	country	year	x1	x2
1	1	A	1995	.	.
2	2	A	1996	.	.
3	3	A	1997	8000.01	6.83
4	4	A	1998	8212.9	2.66
5	5	A	1999	7847.36	-4.45
6	6	A	2000	7702.89	-1.84
7	7	A	2001	7288.48	-5.38
8	8	A	2002	6430.98	-11.77
9	9	A	2003	6932.45	7.8
10	10	A	2004	7486.24	7.99
11	11	A	2005	8094.17	8.12
12	12	B	1995	18268.01	2.87
13	13	B	1996	18738.99	2.58
14	14	B	1997	19360.46	3.32
15	15	B	1998	20151.42	4.09
16	16	B	1999	20715.54	2.8
17	17	B	2000	20866.9	.73
18	18	B	2001	21364.02	2.38
19	19	B	2002	21801.41	2.05
20	20	B	2003	22404.59	2.77
21	21	B	2004	22676.26	1.21

Reshape long to wide (Stata, 1)

You want to go from…

id	time	r
1	1	2
1	2	7
1	3	1
2	1	3
2	2	5
2	3	9
3	1	1
3	2	1
3	3	8

```
reshape wide r, i(id) j(time)
```

to…

id	r.time1	r.time2	r.time3
1	2	7	1
2	3	5	9
3	1	1	8

EXAMPLE: If you have a dataset like this one (click here to get it), we need to change the date variable as follows:

```
tostring month year, replace
gen date=year+"_0"+month if length(month)==1
replace date=year+"_"+month if date==""
drop year month
order id date
```

	id	year	month	return	interest
1	105.1	2002	11	1.307071	.87494
2	105.1	2002	12	1.403008	1.019082
3	105.1	2003	1	1.570926	1.152942
4	105.1	2003	2	1.894784	1.307366
5	105.1	2003	3	1.798847	1.235295
6	105.1	2003	4	1.7628	1.173506
7	105.1	2003	5	2.026655	1.297084
8	105.1	2003	6	2.302488	1.708849
9	105.1	2003	7	2.968058	1.749977
10	105.1	2003	8	3.027948	2.161742
11	105.1	2003	9	3.117896	2.238954
12	105.1	2003	10	5.036636	2.753636
13	105.1	2003	11	5.000024	3.542246
14	105.1	2003	12	7.469865	4.266157
15	105.1	2004	1	8.072268	5.145268
16	105.1	2004	2	7.95181	5.015967
17	105.1	2004	3	8.192726	5.843377
18	105.1	2004	4	8.493984	4.395458
19	105.1	2004	5	5.843343	3.542246
20	105.1	2004	6	5.458126	3.602774
21	105.1	2004	7	5.456205	3.911331

Reshape long to wide (Stata, 2)

The data will look like…

	id	date	return	interest
1	105.1	2002_11	1.307071	.87494
2	105.1	2002_12	1.403008	1.019082
3	105.1	2003_01	1.570926	1.152942
4	105.1	2003_02	1.894784	1.307366
5	105.1	2003_03	1.798847	1.235295
6	105.1	2003_04	1.7628	1.173506
7	105.1	2003_05	2.026655	1.297084
8	105.1	2003_06	2.302488	1.708849
9	105.1	2003_07	2.968058	1.749977
10	105.1	2003_08	3.027948	2.161742
11	105.1	2003_09	3.117896	2.238954
12	105.1	2003_10	5.036636	2.753636
13	105.1	2003_11	5.000024	3.542246
14	105.1	2003_12	7.469865	4.266157
15	105.1	2004_01	8.072268	5.145268
16	105.1	2004_02	7.95181	5.015967
17	105.1	2004_03	8.192726	5.843377
18	105.1	2004_04	8.493984	4.395458
19	105.1	2004_05	5.843343	3.542246
20	105.1	2004_06	5.458126	3.602774
21	105.1	2004_07	5.456205	3.911331

To reshape type

```
reshape wide return interest, i(id) j(date) str
```

```
. reshape wide return interest, i(id) j(date) str
(note: j = 1998_11 1998_12 1999_01 1999_02 1999_03 1999_04 1999_05 1999_06 1999_07 1999_08 1999_09 1999_10 1999_11
> 1999_12 2000_01 2000_02 2000_03 2000_04 2000_05 2000_06 2000_07 2000_08 2000_09 2000_10 2000_11 2000_12 2001_01 2
> 001_02 2001_03 2001_04 2001_05 2001_06 2001_07 2001_08 2001_09 2001_10 2001_11 2001_12 2002_01 2002_02 2002_03 20
> 02_04 2002_05 2002_06 2002_07 2002_08 2002_09 2002_10 2002_11 2002_12 2003_01 2003_02 2003_03 2003_04 2003_05 200
> 3_06 2003_07 2003_08 2003_09 2003_10 2003_11 2003_12 2004_01 2004_02 2004_03 2004_04 2004_05 2004_06 2004_07 2004
> _08 2004_09 2004_10 2004_11 2004_12 2005_01 2005_02 2005_03 2005_04 2005_05 2005_06 2005_07 2005_08 2005_09 2005_
> 10 2005_11 2005_12 2007_01 2007_02 2007_03 2007_04 2007_05 2007_06 2007_07 2007_08 2007_09 2007_10 2007_11)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                      802   ->     25
Number of variables                   4   ->    195
j variable (97 values)             date   ->   (dropped)
xij variables:
                                 return   ->   return1998_11 return1998_12 ... return2007_11
                               interest   ->   interest1998_11 interest1998_12 ... interest2007_11
-----------------------------------------------------------------------------
```

Where:

wide – Indicates the type of reshape, in this case from long to wide format.
return interest – The variables of interest from long to wide are “return” and “interest” (prefix for the new variables).
i(id) – A unique identifier for the wide format is in variable “id”.
j(date) – Indicates the suffix of “return” and “interest” taken from ”date” (notice “xij” variables:” above)

Reshape long to wide (Stata, 3)

The variable window and the data will look like

Variables

Name	Label
id	
return1998_11	1998_11 return
interest1998_11	1998_11 interest
return1998_12	1998_12 return
interest1998_12	1998_12 interest
return1999_01	1999_01 return
interest1999_01	1999_01 interest
return1999_02	1999_02 return
interest1999_02	1999_02 interest
return1999_03	1999_03 return
interest1999_03	1999_03 interest
return1999_04	1999_04 return
interest1999_04	1999_04 interest
return1999_05	1999_05 return
interest1999_05	1999_05 interest

	id	return1998~1	interes~8_11	return1998~2	interes~8_12	return199~01
1	105.1	.	.	.	.	.
2	121.1	3.4126	2.592616	3.108856	2.331589	3.139705
3	143.1	.	.	.	.	.
4	161.2	.	.	.	.	.
5	162.1	.	.	.	.	.
6	162.2	.	.	.	.	.
7	167.1	19.20548	15.14606	18.16995	13.73898	18.71529

If you want to sort all returns and interest together, run the following commands:

```
xpose, clear varname
sort _varname
xpose, clear
order id
```

Variables

Name	Label
id	
interest1998_11	
interest1998_12	
interest1999_01	
interest1999_02	
interest1999_03	
interest1999_04	
interest1999_05	
interest1999_06	
interest1999_07	
interest1999_08	
interest1999_09	
interest1999_10	
interest1999_11	
interest1999_12	

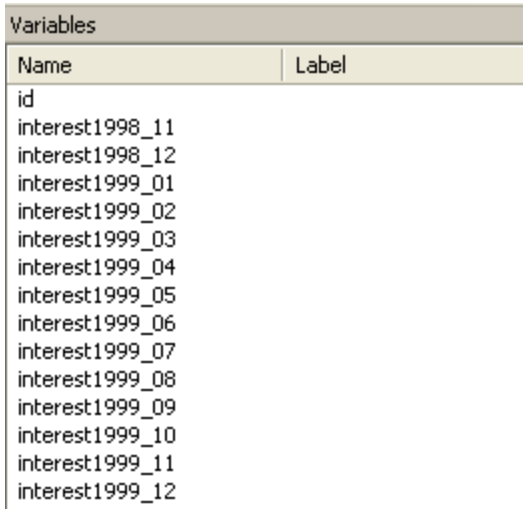
Renaming variables (using `renvars`)

You can use the command `renvars` to shorten the names of the variables…

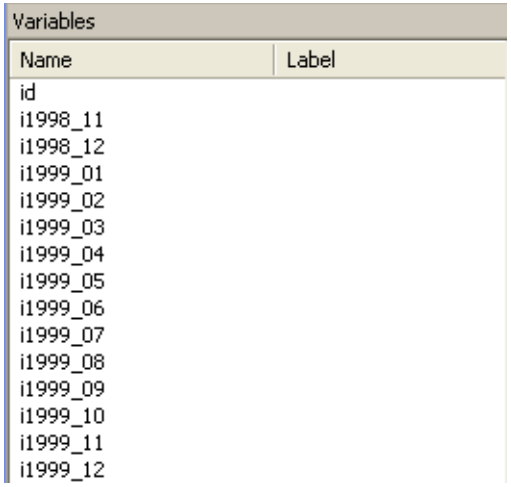
```
renvars  interest1998_11-interest2007_11,  presub(interest i)

renvars  return1998_11-return2007_11,  presub(return r)
```

Before



After



NOTE: You may have to install `renvars` by typing:

```
ssc install renvars
```

Type `help renvars` for more info. Also `help rename`

Descriptive statistics (definitions)

Descriptive statistics are a collection of measurements of two things: *location* and *variability*.

Location tells you the central value of your variable (the mean is the most common measure).

Variability refers to the spread of the data from the center value (i.e. variance, standard deviation).

Statistics is basically the study of what causes variability in the data.

Location	**Variability**
Mean	Variance
Mode	Standard deviation
Median	Range

Descriptive statistics (location)…

Indicator	Definition	Formula	In Excel	In Stata	In R
Location					
Mean	The mean is the sum of the observations divided by the total number of observations. It is the most common indicator of central tendency of a variable	$\overline{X} = \frac{\sum X_i}{n}$	=AVERAGE(range of cells) For example: =AVERAGE(J2:J31)	-tabstat var1, s(mean) or - sum var1	summary(x) mean(x) sapply(x, mean, na.rm=T)
Median	The median is another measure of central tendency. To get the median you have to order the data from lowest to highest. The median is the number in the middle. If the number of cases is odd the median is the single value, for an even number of cases the median is the average of the two numbers in the middle. It is not affected by outliers. Also known as the 50th percentile. 2 6 **7** 8 9 2 6 **7 8** 9 10		=MEDIAN(range of cells)	- tabstat var1, s(median) or - sum var1, detail	summary(x) median(x) sapply(x, median, na.rm=T) #median
Mode	The mode refers to the most frequent, repeated or common number in the data		=MODE(range of cells)	mmodes var1	table(x) (frequency table)

NOTE: For `mmodes` you may have to install it by `typing ssc install mmodes`. You can estimate all statistics in Excell using "Descriptive Statistics" in "Analysis Toolpack". In Stata by typing all statistics in the parenthesis `tabstat var1, s(mean median)`. In R see http://www.ats.ucla.edu/stat/r/faq/basic_desc.htm

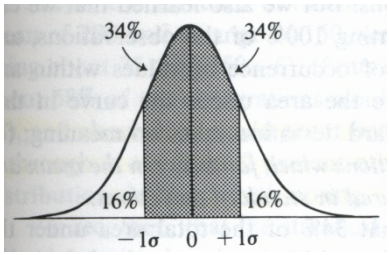
Descriptive statistics (variability)…

Indicator	Definition	Formula	In Excel	In Stata	In R
Variability					
Variance	The variance measures the dispersion of the data from the mean. It is the simple mean of the squared distance from the mean.	$s^2 = \frac{\sum (X_i - \bar{X})^2}{(n-1)}$	=VAR(range of cells)	- tabstat var1, s(variance) or - sum var1, detail	var(x) sapply(x, var, na.rm=T)
Standard deviation	The standard deviation is the squared root of the variance. Indicates how close the data is to the mean. Assuming a normal distribution: • 68% of the values are within 1 sd (.99) • 95% within 2 sd (1.96) • 99% within 3 sd (2.58).	$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{(n-1)}}$	=STDEV(range of cells)	- tabstat var1, s(sd) or - sum var1, detail	sd(x) sapply(x, sd, na.rm=T)
Range	Range is a measure of dispersion. It is simple the difference between the largest and smallest value, “max” – “min”.		=MAX(range of cells) - MIN(same range of cells)	tabstat var1, s(range)	range=(max(x)-min(x));range

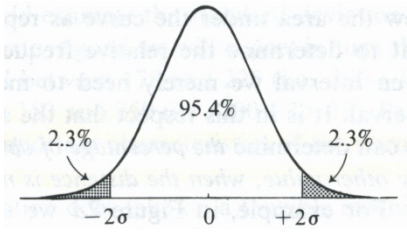
NOTE: You can estimate all statistics in Excell using “Descriptive Statistics” in “Analysis Toolpack”. In Stata by typing all statistics in the parenthesis `tabstat var1, s(mean median variance sd range)`. In R see http://www.ats.ucla.edu/stat/r/faq/basic_desc.htm

Descriptive statistics (standard deviation)

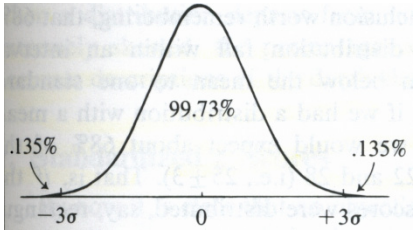
1sd



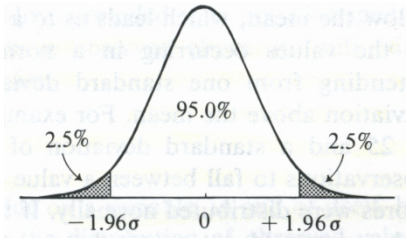
2sd



3sd



1.96sd



Source: Kachigan, Sam K., *Statistical Analysis. An Interdisciplinary Introduction to Univariate & Multivariate Methods*, 1986, p.61

Descriptive statistics (z-scores)…

z-scores show how many standard deviations a single value is from the mean. Having the mean is not enough.

$$z = \frac{x_i - \mu}{\sigma}$$

Student	x_i	Mean SAT score	sd	z-score	% (below)	%(above)
A	1842	1849	275	-0.03	49.0%	51.0%
B	1907	1849	275	0.21	58.4%	41.6%
C	2279	1849	275	1.56	94.1%	5.9%

Student	x_i	Mean SAT score	sd	z-score	% (below)	%(above)
A	1842	1849	162	-0.04	48.3%	51.7%
B	1907	1849	162	0.36	64.0%	36.0%
C	2279	1849	162	2.65	99.6%	0.4%

Student	x_i	Mean SAT score	sd	z-score	% (below)	%(above)
A	1855	1858	162	-0.02	49.3%	50.7%
B	1917	1858	162	0.36	64.2%	35.8%
C	2221	1858	162	2.24	98.7%	1.3%

NOTE: To get the %(below) you can use the tables at the end of any statistics book or in Excel use `=normsdist(z-score)`. %(above) is just `1-%(below)`.

In Stata type:

```
egen z_var1=std(var1)
gen below=normal(z_var1)
gen above=1-below
```

Descriptive statistics (distribution)…

Indicator	Definition	Formula	In Excel	In Stata	In R
Variability					
Standard error (deviation) of the mean	Indicates how close the sample mean is from the 'true' population mean. It increases as the variation increases and it decreases as the sample size goes up. It provides a measure of uncertainty.	$SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$	=(STDEV(range of cells))/(SQRT(COUNT(same range of cells))).	tabstat var1, s(semean)	sem=sd(x)/sqrt (length(x)); sem
Confidence intervals for the mean	The range where the 'true' value of the mean is likely to fall most of the time	$CI_{\bar{X}} = \bar{X} \pm SE_{\bar{X}} * Z$	Use "Descriptive Statistics" in the "Data Analysis" tab (1)	ci var1	Use package "pastecs"
Distribution					
Skewness	Measures the symmetry of the distribution (whether the mean is at the center of the distribution). The skewness value of a normal distribution is 0. A negative value indicates a skew to the left (left tail is longer that the right tail) and a positive values indicates a skew to the right (right tail is longer than the left one)	$Sk = \frac{\sum (X_i - \bar{X})^3}{(n-1)s^3}$	=SKEW(range of cells)	-tabstat var1, s(skew) - sum var1, detail	Custom estimation
Kurtosis	Measures the peakedness (or flatness) of a distribution. A normal distribution has a value of 3. A kurtosis >3 indicates a sharp peak with heavy tails closer to the mean (leptokurtic). A kurtosis < 3 indicates the opposite a flat top (platykurtic).	$K = \frac{\sum (X_i - \bar{X})^4}{(n-1)s^4}$	=KURT(range of cells)	-tabstat var1, s(k) - sum var1, detail	Custom estimation kurtosis(x)

Notation:

X_i = individual value of X
X(bar) = mean of X
n = sample size
s^2 = variance
s = standard deviation
$SE_{X(bar)}$ = standard error of the mean
Z = critical value (Z=1.96 give a 95% certainty)

For more info check the module "Descriptive Statistics with Excel/Stata" in http://dss.princeton.edu/training/

(1) For Excel 2007 http://office.microsoft.com/en-us/excel/HP100215691033.aspx
For Excel 2003 http://office.microsoft.com/en-us/excel/HP011277241033.aspx

Confidence intervals…

Confidence intervals are ranges where the true mean is expected to lie.

Student	x_i	Mean SAT score	sd	N	SE	Lower(95%)	Upper(95%)
A	1842	1849	275	30	50	1751	1947
B	1907	1849	275	30	50	1751	1947
C	2279	1849	275	30	50	1751	1947

Student	x_i	Mean SAT score	sd	N	SE	Lower(95%)	Upper(95%)
A	1842	1849	162	30	30	1791	1907
B	1907	1849	162	30	30	1791	1907
C	2279	1849	162	30	30	1791	1907

Student	x_i	Mean SAT score	sd	N	SE	Lower(95%)	Upper(95%)
A	1855	1858	162	30	30	1800	1916
B	1917	1858	162	30	30	1800	1916
C	2221	1858	162	30	30	1800	1916

lower(95%) = (Mean SAT score) – (SE*1.96)
upper(95%) = (Mean SAT score) + (SE*1.96)

Coefficient of variation (CV)…

Measure of dispersion, helps compare variation across variables with different units. A variable with higher coefficient of variation is more dispersed than one with lower CV.

	A	**B**	**B/A**
	Mean	**Standard Deviation**	**Coefficient of variation**
Age (years)	25	6.87	27%
SAT	1849	275.11	15%
Average score (grade)	80	10.11	13%
Height (in)	66	4.66	7%
Newspaper readership (times/wk)	5	1.28	26%

CV works only with variables with positive values.

Examples (Excel)

Click here to get the table

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)	Height (in)	Newspaper readership (times/wk)
2	1	DOE01	JANE01	Los Angeles	California	Female	Graduate	Politics	US	30	2263	67	61	5
3	2	DOE02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63	64	7
4	3	DOE01	JOE01	Elmira	New York	Male	Graduate	Math	US	26	2221	78	73	6
5	4	DOE02	JOE02	Lackawana	New York	Male	Graduate	Econ	US	33	1716	78	68	3
6	5	DOE03	JOE03	Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65	71	6
7	6	DOE04	JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69	67	5
8	7	DOE05	JOE05	Cimax	North Carolina	Male	Graduate	Politics	US	39	1577	96	70	5
9	8	DOE03	JANE03	Liberal	Kansas	Female	Undergraduate	Politics	US	21	1842	87	62	5
10	9	DOE04	JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada	18	1813	91	62	6
11	10	DOE05	JANE05	New York	New York	Female	Graduate	Math	US	33	2041	71	66	5
12	11	DOE06	JOE06	Hot Coffe	Mississippi	Male	Undergraduate	Econ	US	18	1787	82	67	3
13	12	DOE06	JANE06	Java	Virginia	Female	Graduate	Math	US	38	1513	79	59	5
14	13	DOE07	JOE07	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	30	1637	79	63	4
15	14	DOE08	JOE08	Moscow	Russia	Male	Graduate	Politics	Russia	30	1512	70	75	6
16	15	DOE07	JANE07	Drunkard Creek	New York	Female	Undergraduate	Math	US	21	1338	82	64	5
17	16	DOE08	JANE08	Mexican Hat	Utah	Female	Undergraduate	Econ	US	18	1821	80	63	3
18	17	DOE09	JANE09	Amsterdam	Holland	Female	Undergraduate	Math	Holland	19	1494	75	60	3
19	18	DOE10	JANE10	Mexico	Mexico	Female	Graduate	Politics	Mexico	31	2248	95	59	4
20	19	DOE11	JANE11	Caracas	Venezuela	Female	Undergraduate	Math	Venezuela	18	2252	92	68	5
21	20	DOE09	JOE09	San Juan	Puerto Rico	Male	Graduate	Politics	US	33	1923	95	63	7
22	21	DOE12	JANE12	Remote	Oregon	Female	Undergraduate	Econ	US	19	1727	67	62	7
23	22	DOE10	JOE10	New York	New York	Male	Undergraduate	Econ	US	21	1872	82	73	4
24	23	DOE13	JANE13	The X	Massachusetts	Female	Graduate	Politics	US	25	1767	89	68	6
25	24	DOE14	JANE14	Beijing	China	Female	Undergraduate	Math	China	18	1643	79	65	6
26	25	DOE11	JOE11	Stockholm	Sweden	Male	Undergraduate	Politics	Sweden	19	1919	88	64	4
27	26	DOE12	JOE12	Embarrass	Minnesota	Male	Graduate	Econ	US	28	1434	96	71	4
28	27	DOE13	JOE13	Intercourse	Pennsylvania	Male	Undergraduate	Math	US	20	2119	88	71	5
29	28	DOE15	JANE15	Loco	Oklahoma	Female	Undergraduate	Econ	US	20	2309	64	68	6
30	29	DOE14	JOE14	Buenos Aires	Argentina	Male	Graduate	Politics	Argentina	30	2279	85	72	3
31	30	DOE15	JOE15	Acme	Louisiana	Male	Undergraduate	Econ	US	19	1907	79	74	3

Use “Descriptive Statistics” in the “Data Analysis” tab.

Age		*SAT*		*Average score (grade)*		*Height (in)*		*Newspaper readership (times/wk)*	
Mean	25.2	Mean	1848.9	Mean	80.40091482	Mean	66.43333333	Mean	4.866666667
Standard Error	1.254325848	Standard Error	50.22838301	Standard Error	1.845084499	Standard Error	0.850535103	Standard Error	0.233579509
Median	23	Median	1817	Median	79.74967997	Median	66.5	Median	5
Mode	19	Mode	#N/A	Mode	67	Mode	68	Mode	5
Standard Deviation	6.870225615	Standard Deviation	275.112184	Standard Deviation	10.10594401	Standard Deviation	4.658572619	Standard Deviation	1.27936766
Sample Variance	47.2	Sample Variance	75686.71379	Sample Variance	102.1301043	Sample Variance	21.70229885	Sample Variance	1.636781609
Kurtosis	-1.049751548	Kurtosis	-0.846633469	Kurtosis	-0.991907645	Kurtosis	-1.066828463	Kurtosis	-0.972412281
Skewness	0.557190515	Skewness	0.155667999	Skewness	-0.112360607	Skewness	0.171892733	Skewness	-0.051910426
Range	21	Range	971	Range	32.88251459	Range	16	Range	4
Minimum	18	Minimum	1338	Minimum	63	Minimum	59	Minimum	3
Maximum	39	Maximum	2309	Maximum	95.88251459	Maximum	75	Maximum	7
Sum	756	Sum	55467	Sum	2412.027445	Sum	1993	Sum	146
Count	30	Count	30	Count	30	Count	30	Count	30

For Excel 2007 http://office.microsoft.com/en-us/excel/HP100215691033.aspx
For Excel 2003 http://office.microsoft.com/en-us/excel/HP011277241033.aspx

Examples (Stata)

Click here to get the table

	id	lastname	firstname	city	state	gender	studentstatus	major	country	age	sat	averagesco~e	heightin	newspaperr~k
1	1	DOE01	JANE01	Los Angeles	California	Female	Graduate	Politics	US	30	2263	67	61	5
2	2	DOE02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63	64	7
3	3	DOE01	JOE01	Elmira	New York	Male	Graduate	Math	US	26	2221	78	73	6
4	4	DOE02	JOE02	Lackawana	New York	Male	Graduate	Econ	US	33	1716	78	68	3
5	5	DOE03	JOE03	Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65	71	6
6	6	DOE04	JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69	67	5
7	7	DOE05	JOE05	Cimax	North Carolina	Male	Graduate	Politics	US	39	1577	96	70	5
8	8	DOE03	JANE03	Liberal	Kansas	Female	Undergraduate	Politics	US	21	1842	87	62	5
9	9	DOE04	JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada	18	1813	91	62	6
10	10	DOE05	JANE05	New York	New York	Female	Graduate	Math	US	33	2041	71	66	5
11	11	DOE06	JOE06	Hot Coffe	Mississippi	Male	Undergraduate	Econ	US	18	1787	82	67	3
12	12	DOE06	JANE06	Java	Virginia	Female	Graduate	Math	US	38	1513	79	59	5
13	13	DOE07	JOE07	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	30	1637	79	63	4
14	14	DOE08	JOE08	Moscow	Russia	Male	Graduate	Politics	Russia	30	1512	70	75	6
15	15	DOE07	JANE07	Drunkard Creek	New York	Female	Undergraduate	Math	US	21	1338	82	64	5
16	16	DOE08	JANE08	Mexican Hat	Utah	Female	Undergraduate	Econ	US	18	1821	80	63	3
17	17	DOE09	JANE09	Amsterdam	Holland	Female	Undergraduate	Math	Holland	19	1494	75	60	3
18	18	DOE10	JANE10	Mexico	Mexico	Female	Graduate	Politics	Mexico	31	2248	95	59	4
19	19	DOE11	JANE11	Caracas	Venezuela	Female	Undergraduate	Math	Venezuela	18	2252	92	68	5
20	20	DOE09	JOE09	San Juan	Puerto Rico	Male	Graduate	Politics	US	33	1923	95	63	7
21	21	DOE12	JANE12	Remote	Oregon	Female	Undergraduate	Econ	US	19	1727	67	62	7
22	22	DOE10	JOE10	New York	New York	Male	Undergraduate	Econ	US	21	1872	82	73	4
23	23	DOE13	JANE13	The X	Massachusetts	Female	Graduate	Politics	US	25	1767	89	68	6
24	24	DOE14	JANE14	Beijing	China	Female	Undergraduate	Math	China	18	1643	79	65	6
25	25	DOE11	JOE11	Stockholm	Sweden	Male	Undergraduate	Politics	Sweden	19	1919	88	64	4
26	26	DOE12	JOE12	Embarrass	Minnesota	Male	Graduate	Econ	US	28	1434	96	71	4
27	27	DOE13	JOE13	Intercourse	Pennsylvania	Male	Undergraduate	Math	US	20	2119	88	71	5
28	28	DOE15	JANE15	Loco	Oklahoma	Female	Undergraduate	Econ	US	20	2309	64	68	6
29	29	DOE14	JOE14	Buenos Aires	Argentina	Male	Graduate	Politics	Argentina	30	2279	85	72	3
30	30	DOE15	JOE15	Acme	Louisiana	Male	Undergraduate	Econ	US	19	1907	79	74	3

```
rename  averagescoregrade score
rename  newspaperreadershiptimeswk read


tabstat  age sat score heightin read,
s(mean semean median sd var skew k
count sum range min max )
```

stats	age	sat	score	heightin	read
mean	25.2	1848.9	80.36667	66.43333	4.866667
se(mean)	1.254326	50.22838	1.846079	.8505351	.2335795
p50	23	1817	79.5	66.5	5
sd	6.870226	275.1122	10.11139	4.658573	1.279368
variance	47.2	75686.71	102.2402	21.7023	1.636782
skewness	.5289348	.1477739	-.1017756	.1631759	-.049278
kurtosis	1.923679	2.094488	1.966325	1.909319	1.988717
N	30	30	30	30	30
sum	756	55467	2411	1993	146
range	21	971	33	16	4
min	18	1338	63	59	3
max	39	2309	96	75	7

Examples (R)

```
students<-read.table("H:/students.csv", sep=",",header=T)
```

```
> students
   ID  last  first           City          State Gender        status     Major   Country Age  SAT score height read
1   1 DOE01 JANE01    Los Angeles     California Female      Graduate Politics        US  30 2263    67     61    5
2   2 DOE02 JANE02         Sedona        Arizona Female Undergraduate      Math        US  19 2006    63     64    7
3   3 DOE01  JOE01         Elmira       New York   Male      Graduate      Math        US  26 2221    78     73    6
4   4 DOE02  JOE02      Lackawana       New York   Male      Graduate      Econ        US  33 1716    78     68    3
5   5 DOE03  JOE03       Defiance           Ohio   Male      Graduate      Econ        US  37 1701    65     71    6
6   6 DOE04  JOE04       Tel Aviv         Israel   Male      Graduate      Econ    Israel  25 1786    69     67    5
7   7 DOE05  JOE05          Cimax North Carolina   Male      Graduate Politics        US  39 1577    96     70    5
8   8 DOE03 JANE03        Liberal         Kansas Female Undergraduate Politics        US  21 1842    87     62    5
9   9 DOE04 JANE04       Montreal         Canada Female Undergraduate      Math    Canada  18 1813    91     62    6
10 10 DOE05 JANE05       New York       New York Female      Graduate      Math        US  33 2041    71     66    5
11 11 DOE06  JOE06      Hot Coffe    Mississippi   Male Undergraduate      Econ        US  18 1787    82     67    3
12 12 DOE06 JANE06           Java       Virginia Female      Graduate      Math        US  38 1513    79     59    5
13 13 DOE07  JOE07         Varna       Bulgaria   Male      Graduate Politics  Bulgaria  30 1637    79     63    4
14 14 DOE08  JOE08         Moscow         Russia   Male      Graduate Politics    Russia  30 1512    70     75    6
15 15 DOE07 JANE07 Drunkard Creek       New York Female Undergraduate      Math        US  21 1338    82     64    5
16 16 DOE08 JANE08    Mexican Hat           Utah Female Undergraduate      Econ        US  18 1821    80     63    3
17 17 DOE09 JANE09      Amsterdam        Holland Female Undergraduate      Math   Holland  19 1494    75     60    3
18 18 DOE10 JANE10         Mexico         Mexico Female      Graduate Politics    Mexico  31 2248    95     59    4
19 19 DOE11 JANE11        Caracas      Venezuela Female Undergraduate      Math Venezuela  18 2252    92     68    5
20 20 DOE09  JOE09       San Juan    Puerto Rico   Male      Graduate Politics        US  33 1923    95     63    7
21 21 DOE12 JANE12         Remote         Oregon Female Undergraduate      Econ        US  19 1727    67     62    7
22 22 DOE10  JOE10       New York       New York   Male Undergraduate      Econ        US  21 1872    82     73    4
23 23 DOE13 JANE13          The X Massachusetts  Female      Graduate Politics        US  25 1767    89     68    6
24 24 DOE14 JANE14        Beijing          China Female Undergraduate      Math     China  18 1643    79     65    6
25 25 DOE11  JOE11      Stockholm         Sweden   Male Undergraduate Politics    Sweden  19 1919    88     64    4
26 26 DOE12  JOE12      Embarrass      Minnesota   Male      Graduate      Econ        US  28 1434    96     71    4
27 27 DOE13  JOE13    Intercourse   Pennsylvania   Male Undergraduate      Math        US  20 2119    88     71    5
28 28 DOE15 JANE15           Loco       Oklahoma Female Undergraduate      Econ        US  20 2309    64     68    6
29 29 DOE14  JOE14   Buenos Aires      Argentina   Male      Graduate Politics Argentina  30 2279    85     72    3
30 30 DOE15  JOE15          Acme      Louisiana   Male Undergraduate      Econ        US  19 1907    79     74    3
> 
```

```
install.packages("pastecs")
```

```
> library(pastecs)
Loading required package: boot
> stat.desc(students[10:14])
                    Age          SAT        score       height         read
nbr.val       30.000000 3.000000e+01   30.0000000 3.000000e+01 30.0000000
nbr.null       0.000000 0.000000e+00    0.0000000 0.000000e+00  0.0000000
nbr.na         0.000000 0.000000e+00    0.0000000 0.000000e+00  0.0000000
min           18.000000 1.338000e+03   63.0000000 5.900000e+01  3.0000000
max           39.000000 2.309000e+03   96.0000000 7.500000e+01  7.0000000
range         21.000000 9.710000e+02   33.0000000 1.600000e+01  4.0000000
sum          756.000000 5.546700e+04 2411.0000000 1.993000e+03 146.0000000
median        23.000000 1.817000e+03   79.5000000 6.650000e+01  5.0000000
mean          25.200000 1.848900e+03   80.3666667 6.643333e+01  4.8666667
SE.mean        1.254326 5.022838e+01    1.8460790 8.505351e-01  0.2335795
CI.mean.0.95   2.565384 1.027286e+02    3.7756555 1.739540e+00  0.4777237
var           47.200000 7.568671e+04  102.2402299 2.170230e+01  1.6367816
std.dev        6.870226 2.751122e+02   10.1113911 4.658573e+00  1.2793677
coef.var       0.272628 1.487978e-01    0.1258157 7.012402e-02  0.2628838
> 
```


Source: http://www.ats.ucla.edu/stat/r/faq/basic_desc.htm

Useful links / Recommended books/References

- DSS Online Training Section http://dss.princeton.edu/training/
- UCLA Resources http://www.ats.ucla.edu/stat/
- DSS help-sheets for STATA http://dss/online_help/stats_packages/stata/stata.htm
- *Introduction to Stata* (PDF), Christopher F. Baum, Boston College, USA. "A 67-page description of Stata, its key features and benefits, and other useful information." http://fmwww.bc.edu/GStat/docs/StataIntro.pdf
- STATA FAQ website http://stata.com/support/faqs/
- Princeton DSS Libguides http://libguides.princeton.edu/dss

Books

- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Data analysis using regression and multilevel/hierarchical models* / Andrew Gelman, Jennifer Hill. Cambridge ; New York : Cambridge University Press, 2007.
- *Econometric analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *Designing Social Inquiry: Scientific Inference in Qualitative Research* / *Gary King, Robert O. Keohane, Sidney Verba, Princeton University Press*, 1994.
- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* / Gary King, Cambridge University Press, 1989
- *Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods /* Sam Kachigan, New York : Radius Press, c1986
- *Statistics with Stata (updated for version 9) /* Lawrence Hamilton, Thomson Books/Cole, 2006