Getting Started in Frequencies, Crosstab, Factor and Regression Analysis

(v. 2.0)

Oscar Torres-Reyna

otorres@princeton.edu

August 2008

http://www.princeton.edu/~otorres/

Case study: intro

Search here in the home page for this dataset

Metadata

Search Results

Title	**Time Magazine/Abt SRBI Poll # 2008-4567: America by the Numbers** [Study# USSRBI2008-4567]
Survey Firm	Abt SRBI, Inc. (Schulman, Ronca, & Bucuvalas, Inc.)
Survey Sponsor	Time Magazine
Field Dates	October 3-6, 2008
Sample	Registered likely voters
Sample Size	1,053
Sample Notes	Respondents were interviewed via landlines and cellular telephones.
Variables	136

Major Topics Covered

Voter history/intent (3); 2008 presidential election (1); Obama/Biden vs. McCain/Palin (4); rating political leaders/people (10); Sarah Palin vs. Joe Biden (5); Barack Obama vs. John McCain (3); source of news (15); George W. Bush job performance (1); direction of country (1); economy (2); social contract (2); comparing now and then (1); mortgage recovery plan (1); opinion on certain statements (10); war in Iraq (1); people in the news (5).

Codebook in two formats

Documentation Download

Study documentation files are available for free download.
PDF (186kb)
Word (440kb)

Datasets, two formats: ACII and SPSS

RoperExpress

The following files are available only to **RoperExpress** Users and Members.

Data Sets
ASCII (385kb)
SPSS portable (496kb)

Study File Listing and Other Notes
Text (3kb)

Marginals

Data tables/Frequencies
PDF (54kb)

NOTE: When data is not available in Stata, you can download the SPSS portable (*.por), open it using SPSS (available at the DSS lab) and saving it as Stata.

Case study: frequencies

Distribution of electoral preferences and gender. According to the codebook ‘q5’ has the electoral question and ‘qa’ gender.

```
. tab q5  /*No weights*/
```

Q5. If the Presidential election were held today and the candidates were Barack	Freq.	Percent	Cum.
Barack Obama and Joe Biden, the Democra	481	45.68	45.68
John McCain and Sarah Palin, the Republ	464	44.06	89.74
(VOL) Other/Neither	21	1.99	91.74
(VOL) Undecided/Don't know/no answer	87	8.26	100.00
Total	1,053	100.00	

No weights

```
. tab q5 [aweight=weight] /*With weights*/
```

Q5. If the Presidential election were held today and the candidates were Barack	Freq.	Percent	Cum.
Barack Obama and Joe Biden, the Democra	504.337749	47.90	47.90
John McCain and Sarah Palin, the Republ	449.487545	42.69	90.58
(VOL) Other/Neither	20.5570831	1.95	92.53
(VOL) Undecided/Don't know/no answer	78.61762284	7.47	100.00
Total	1,053	100.00	

Using weights

```
. tab qa  /*No weights*/
```

A. Gender (DO NOT ASK)	Freq.	Percent	Cum.
Male	493	46.82	46.82
Female	560	53.18	100.00
Total	1,053	100.00	

No weights

```
. tab qa [aweight=weight] /*With weights*/
```

A. Gender (DO NOT ASK)	Freq.	Percent	Cum.
Male	500.388396	47.52	47.52
Female	552.611604	52.48	100.00
Total	1,053	100.00	

Using weights

NOTE: At this point, it is strongly recommended to open a log to keep a record of your work and to extract output, type:

```
log using mywork.log
```

You could also open a do-file by typing `doedit` and copy your commands there.

Case study: Electoral preferences by gender

```
. tab q5 qa [aw=weight], col row /*Electoral preferences by gender*/
```

Key
frequency *row percentage* *column percentage*

Q5. If the Presidential election were held today and the candidates were Barack	A. Gender (DO NOT ASK) Male	Female	Total
Barack Obama and Joe	209.42078 41.52 41.85	294.91697 58.48 53.37	504.33775 100.00 47.90
John McCain and Sarah	252.9313 56.27 50.55	196.55625 43.73 35.57	449.487545 100.00 42.69
(VOL) Other/Neither	10.055739 48.92 2.01	10.5013441 51.08 1.90	20.557083 100.00 1.95
(VOL) Undecided/Don't	27.980574 35.59 5.59	50.637048 64.41 9.16	78.617623 100.00 7.47
Total	500.3884 47.52 100.00	552.6116 52.48 100.00	1,053 100.00 100.00

Case study: Electoral preferences by age

```
. tab q5 f1 [aw=weight], col row /*Electoral preferences by age*/
```

Key
frequency *row percentage* *column percentage*

Q5. If the Presidential election were held today and the candidates were Barack	F1. What is your age? 18-24	25-29	30-34	35-39	40-44	45-54	55-64	F1. What is your age? 65 or old	(VOL) No	Total
Barack Obama and Joe	29.355119	26.435913	40.727272	46.595118	38.610873	129.51971	86.169373	102.39238	4.5319886	504.33775
	5.82	5.24	8.08	9.24	7.66	25.68	17.09	20.30	0.90	100.00
	77.57	50.53	40.92	51.51	37.59	51.68	50.32	43.21	40.00	47.90
John McCain and Sarah	6.2229886	22.18839	54.883049	36.825588	51.046351	99.992283	69.037199	104.76215	4.5295414	449.487545
	1.38	4.94	12.21	8.19	11.36	22.25	15.36	23.31	1.01	100.00
	16.44	42.42	55.14	40.71	49.69	39.90	40.31	44.21	39.98	42.69
(VOL) Other/Neither	0	0	2.1209543	2.4419715	4.51458561	3.1358789	2.7783459	5.56534701	0	20.557083
	0.00	0.00	10.32	11.88	21.96	15.25	13.52	27.07	0.00	100.00
	0.00	0.00	2.13	2.70	4.39	1.25	1.62	2.35	0.00	1.95
(VOL) Undecided/Don't	2.2672181	3.6879373	1.809561	4.5920698	8.5570854	17.952531	13.264407	24.219596	2.2672179	78.617623
	2.88	4.69	2.30	5.84	10.88	22.84	16.87	30.81	2.88	100.00
	5.99	7.05	1.82	5.08	8.33	7.16	7.75	10.22	20.01	7.47
Total	37.845325	52.312241	99.540836	90.454747	102.7289	250.600407	171.24932	236.93948	11.328748	1,053
	3.59	4.97	9.45	8.59	9.76	23.80	16.26	22.50	1.08	100.00
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Case study: Electoral preferences by educational attainment

```
. tab q5 f4 [aw=weight], col row /*Electoral preferences by education*/
```

Key
frequency *row percentage* *column percentage*

Q5. If the Presidential election were held today and the candidates were Barack	F4. What is the highest grade of schooling that you've completed? 8th grade	Some high	High scho	Some coll	College g	Postgradu	(VOL) No	Total
Barack Obama and Joe	2.2991619 0.46 41.27	3.883265 0.77 33.65	81.589679 16.18 44.32	113.53524 22.51 45.42	169.39657 33.59 44.89	130.23545 25.82 59.52	3.3983797 0.67 60.03	504.33775 100.00 47.90
John McCain and Sarah	3.2718681 0.73 58.73	6.1159475 1.36 53.00	76.7484051 17.07 41.69	116.69213 25.96 46.68	170.30303 37.89 45.13	74.093841 16.48 33.86	2.2623235 0.50 39.97	449.487545 100.00 42.69
(VOL) Other/Neither	0 0.00 0.00	0 0.00 0.00	3.7389017 18.19 2.03	3.382658 16.45 1.35	9.8911577 48.12 2.62	3.5443656 17.24 1.62	0 0.00 0.00	20.557083 100.00 1.95
(VOL) Undecided/Don't	0 0.00 0.00	1.5397725 1.96 13.34	22.004128 27.99 11.95	16.367784 20.82 6.55	27.7818421 35.34 7.36	10.924096 13.90 4.99	0 0.00 0.00	78.617623 100.00 7.47
Total	5.57103 0.53 100.00	11.538985 1.10 100.00	184.08111 17.48 100.00	249.97781 23.74 100.00	377.3726 35.84 100.00	218.79775 20.78 100.00	5.6607032 0.54 100.00	1,053 100.00 100.00

Case study: Electoral preferences by income

```
. tab q5 f13 [aw=weight], col row /*Electoral preferences by income*/
```

Key
frequency *row percentage* *column percentage*

Q5. If the Presidential election were held today and the candidates were Barack	F13. Finally, just for classification purposes, was your total family income bef Less than	$20,000 t	$35,000 t	$50,000 t	$75,000 t	$100,000	or $150,0	(VOL) No	Total
Barack Obama and Joe	37.525195	51.14097	72.715849	122.78749	59.632459	69.732723	51.1129092	39.690155	504.33775
	7.44	10.14	14.42	24.35	11.82	13.83	10.13	7.87	100.00
	60.42	49.40	48.59	57.51	39.05	46.16	44.46	37.63	47.90
John McCain and Sarah	18.630762	39.764056	64.4115908	69.827216	86.023642	68.843117	54.640308	47.346852	449.487545
	4.14	8.85	14.33	15.53	19.14	15.32	12.16	10.53	100.00
	30.00	38.41	43.04	32.71	56.34	45.57	47.53	44.88	42.69
(VOL) Other/Neither	1.5060026	.88321203	3.2060684	2.5018142	2.1243815	3.0806277	2.200355	5.0546217	20.557083
	7.33	4.30	15.60	12.17	10.33	14.99	10.70	24.59	100.00
	2.42	0.85	2.14	1.17	1.39	2.04	1.91	4.79	1.95
(VOL) Undecided/Don't	4.4480018	11.739914	9.3136182	18.37691	4.9181423	9.409895	7.01703324	13.3941079	78.617623
	5.66	14.93	11.85	23.38	6.26	11.97	8.93	17.04	100.00
	7.16	11.34	6.22	8.61	3.22	6.23	6.10	12.70	7.47
Total	62.109961	103.52815	149.64713	213.49343	152.69863	151.06636	114.97061	105.48574	1,053
	5.90	9.83	14.21	20.27	14.50	14.35	10.92	10.02	100.00
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Case study: Electoral preferences by employment status

```
. tab q5 f8 [aw=weight], col row /*Electoral preferences by employment status*/
```

Key
frequency *row percentage* *column percentage*

Q5. If the Presidential election were held today and the candidates were Barack	f8 Employed	Employed	Laid off	Retired	Student	Homemaker	Something	(VOL) No	Total
Barack Obama and Joe	263.30095	36.9693237	17.692466	125.50328	15.486465	16.644394	24.6988275	4.0420475	504.33775
	52.21	7.33	3.51	24.88	3.07	3.30	4.90	0.80	100.00
	47.30	52.01	67.23	46.14	82.02	27.25	60.77	64.12	47.90
John McCain and Sarah	252.31686	25.723928	6.1500438	112.5963	1.1268505	37.187532	12.123702	2.2623235	449.487545
	56.13	5.72	1.37	25.05	0.25	8.27	2.70	0.50	100.00
	45.33	36.19	23.37	41.39	5.97	60.89	29.83	35.88	42.69
(VOL) Other/Neither	11.498793	1.6530186	0	6.1126834	0	0	1.2925883	0	20.557083
	55.94	8.04	0.00	29.74	0.00	0.00	6.29	0.00	100.00
	2.07	2.33	0.00	2.25	0.00	0.00	3.18	0.00	1.95
(VOL) Undecided/Don't	29.558151	6.7386098	2.4747578	27.814172	2.2672181	7.2399743	2.52474	0	78.617623
	37.60	8.57	3.15	35.38	2.88	9.21	3.21	0.00	100.00
	5.31	9.48	9.40	10.22	12.01	11.85	6.21	0.00	7.47
Total	556.67476	71.08488	26.3172676	272.02643	18.8805338	61.0719	40.639858	6.304371	1,053
	52.87	6.75	2.50	25.83	1.79	5.80	3.86	0.60	100.00
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Case study: Testing for associations (preparing the data)

Before running any test we need to prepare the data by setting to missing any non-valid response (like “don’t know/no answer/not sure”) unless is relevant to the question. It is important to ‘clean’ the variables for the tests to be as accurate as possible. For demographics we will remove non-response items. Here are a series of commands per variable (columns) to prepare some variables for you to run on your own.

Description	Age	Education	Income	Employment	Gender
creating a new variable	gen age=f1	gen educ=f4	gen income=f13	gen employ=f8	gen gender=qa
exploring the new variable	tab age	tab educ	tab income	tab employ	tab gender
checking for labels from original variable	labelbook f1	labelbook f4	labelbook f13	labelbook f8	labelbook qa
assigning labels to new variable	label value age f1	label value educ f4	label value income f13	label value employ f8	label value gender qa
exploring the new variable	tab age	tab educ	tab income	tab employ	tab gender
setting no response to missing	replace age=. if age>8	replace educ=. if educ==8	replace income=. if income==8	replace employ=. if employ==8	
adding variable labels	label variable age "Age"	label variable educ "Educational attainment"	label variable income "Family income"	label variable employ "Employment status"	
exploring the new variable	tab age	tab educ	tab income	tab employ	

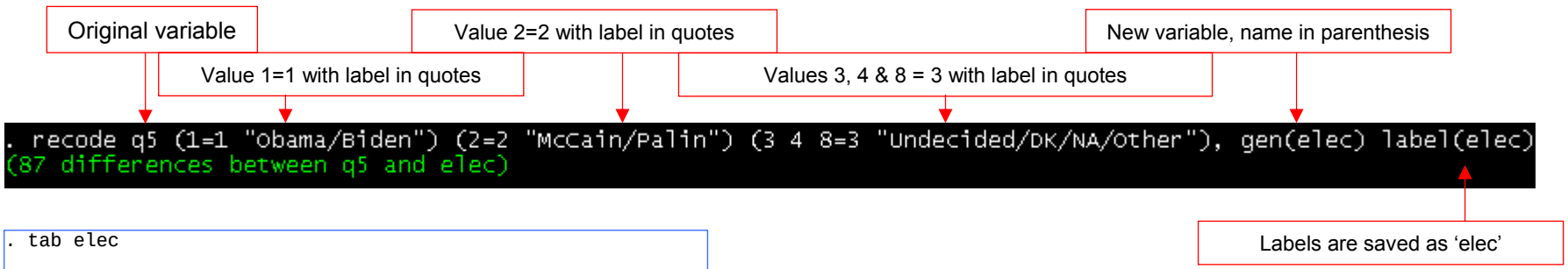
Case study: Testing for associations (preparing the data –cont.)

Here is an easy way to do it by using the command `clonevar` in Stata.

Description	Age	Education	Income	Employment	Gender
creating a new variable	clonevar age=f1	clonevar educ=f4	clonevar income=f13	clonevar employ=f8	clonevar gender=qa
exploring the new variable	tab age	tab educ	tab income	tab employ	tab gender
setting no response to missing	replace age=. if age>8	replace educ=. if educ==8	replace income=. if income==8	replace employ=. if employ==8	
exploring the new variable	tab age	tab educ	tab income	tab employ	

Case study: testing for associations

To find whether there is some association between demographics and electoral preferences we can use chi-square but first we need to 'clean' the electoral variable (q5). Lets create a new variable 'elec' from 'q5'. We will use `recode` for this, type:




```
. tab elec

RECODE of q5 (Q5. If the Presidential election were held today and the candidate | Freq.     Percent        Cum.
------------------------------------------------------------------------------------+-----------------------------------
                                                                       Obama/Biden |        481       45.68       45.68
                                                                      McCain/Palin |        464       44.06       89.74
                                                                 Undecided/DK/NA/O |        108       10.26      100.00
------------------------------------------------------------------------------------+-----------------------------------
                                                                             Total |      1,053      100.00
```

Here is the new variable

```
. tab elec gender, nofreq chi2
          Pearson chi2(2) =  21.0639   Pr = 0.000

. tab elec age, nofreq chi2
          Pearson chi2(14) =  20.6142   Pr = 0.112

. tab elec educ, nofreq chi2
          Pearson chi2(10) =  25.8557   Pr = 0.004

. tab elec income, nofreq chi2
          Pearson chi2(12) =  26.4188   Pr = 0.009

. tab elec employ, nofreq chi2
          Pearson chi2(12) =  21.8394   Pr = 0.039
```

We use the '`nofreq`' option after comma since we are not interested on the crosstabulations but rather on the tests. We can see that gender, education, income and employment status are somehow associated with electoral preferences. Age does not seem to have any association.

Case study: descriptive statistics

When you have continuous data you need to use descriptive statistics. To start exploring this option you can use the summarize command which provides first look at the data (number of observations, mean, standard deviation, minimum and maximum values). Lets take a look at the battery of questions in q8.

```
. summarize  q8a q8b q8c q8d q8e q8f q8g q8h q8i q8j
```

Variable	Obs	Mean	Std. Dev.	Min	Max
q8a	1053	69.26591	117.1873	0	999
q8b	1053	66.09497	108.8857	0	999
q8c	1053	78.26401	146.9197	0	999
q8d	1053	64.73029	124.7597	0	999
q8e	1053	72.20038	173.5736	0	999
q8f	1053	119.5973	237.9549	0	999
q8g	1053	111.4653	223.1448	0	999
q8h	1053	37.36657	60.16443	0	999
q8i	1053	73.7075	137.2898	0	999
q8j	1053	99.63723	189.1918	0	999

The questions ask for answers between 0 and 100. The maximum value 999 represents “Not answer/Not sure” response. The mean and standard deviation factor in the 999 therefore biasing the mean and sd. so we need to set 999 to missing so the values go from 0 to 100.

```
. summarize  x8a x8b x8c x8d x8e x8f x8g x8h x8i x8j
```

Variable	Obs	Mean	Std. Dev.	Min	Max
x8a	1038	55.83044	35.31804	0	100
x8b	1040	54.43365	31.28831	0	100
x8c	1028	55.87257	31.18157	0	100
x8d	1036	49.39961	35.33493	0	100
x8e	1018	40.33595	24.2347	0	100
x8f	982	56.01527	26.50595	0	100
x8g	991	55.93845	22.22173	0	100
x8h	1050	34.61905	31.2718	0	100
x8i	1031	53.96314	23.95454	0	100
x8j	1009	60.41824	22.56533	0	100

Here 999 is set to missing and we have correct statistics (see the slides on ‘preparing the data’ to do this). For presentation purposes we won’t use weights here.

Case study: descriptive statistics

To get more than the mean and sd you can use `tabstat` which offers several options (type help `tabstat` for more details). Notice we use weights here.

In these series of questions ‘0’ means ‘unfavorable’ and ‘100’ favorable.

```
. tabstat  x8a x8b x8c x8d x8e x8f x8g x8h x8i x8j, s(mean median sd var count range min max)
```

stats	x8a	x8b	x8c	x8d	x8e	x8f	x8g	x8h	x8i	x8j
mean	**55.83044**	**54.43365**	**55.87257**	**49.39961**	**40.33595**	**56.01527**	**55.93845**	**34.61905**	**53.96314**	**60.41824**
p50	**60**	**55**	**60**	**50**	**50**	**60**	**55**	**30**	**55**	**65**
sd	**35.31804**	**31.28831**	**31.18157**	**35.33493**	**24.2347**	**26.50595**	**22.22173**	**31.2718**	**23.95454**	**22.56533**
variance	**1247.364**	**978.9581**	**972.2905**	**1248.557**	**587.3207**	**702.5655**	**493.8053**	**977.9253**	**573.82**	**509.194**
N	**1038**	**1040**	**1028**	**1036**	**1018**	**982**	**991**	**1050**	**1031**	**1009**
range	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
min	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
max	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Here is a description of each variable

```
. describe x8*
```

variable name	storage type	display format	value label	variable label
x8a	float	%9.0g		**Obama**
x8b	float	%9.0g		**McCain**
x8c	float	%9.0g		**Biden**
x8d	float	%9.0g		**Palin**
x8e	float	%9.0g		**Congress**
x8f	float	%9.0g		**Congressman**
x8g	float	%9.0g		**Supreme court**
x8h	float	%9.0g		**Pres. Bush**
x8i	float	%9.0g		**State gov**
x8j	float	%9.0g		**Local gov**

Case study: descriptive statistics

Lets explore a combination of commands to get more info out of your data. We will check out the battery of questions in q25

```
. describe q25*

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
q25a            double %10.0g      q25a       Q25. Favor/Oppose: A woman
                                                should be able to get an
                                                abortion if she wants one in
q25b            double %10.0g      q25b       Q25. Favor/Oppose: Gay couples
                                                should be allowed to marry,
                                                giving them full lega
q25c            double %10.0g      q25c       Q25. Favor/Oppose: The
                                                government should provide
                                                health care coverage to all cit
q25d            double %10.0g      q25d       Q25. Favor/Oppose: Government
                                                regulation of financial
                                                institutions should be gre
q25e            double %10.0g      q25e       Q25. Favor/Oppose: The
                                                government should have let
                                                financial institutions that go
q25f            double %10.0g      q25f       Q25. Favor/Oppose: The
                                                government should allow
                                                offshore drilling for oil and
                                                gas
q25g            double %10.0g      q25g       Q25. Favor/Oppose: Congress
                                                should pass stricter laws to
                                                protect the environment
q25h            double %10.0g      q25h       Q25. Favor/Oppose: Our troops
                                                should stay in Iraq without a
                                                timetable for withdr
q25i            double %10.0g      q25i       Q25. Favor/Oppose: Government
                                                should cut taxes on businesses
                                                to help the economy
q25j            double %10.0g      q25j       Q25. Favor/Oppose: The
                                                government should help people
                                                who can't afford their mort
```

```
. sum q25*

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        q25a |       1053    6.907882    10.29471         0         99
        q25b |       1053    5.424501    10.47273         0         99
        q25c |       1053    7.321937    12.96045         0         99
        q25d |       1053    8.025641    11.76441         0         99
        q25e |       1053    9.676163    18.00545         0         99
-------------+---------------------------------------------------------
        q25f |       1053    8.073124    9.473867         0         99
        q25g |       1053    7.635328    11.42399         0         99
        q25h |       1053    6.269706     11.7495         0         99
        q25i |       1053    8.096866    14.78033         0         99
        q25j |       1053    7.317189    14.04718         0         99
```

The questions ask for answers between 0 and 10 (see the codebook) . The maximum value 99 (below) represents “Not answer/Not sure” response.

The mean and standard deviation factor in the 99 therefore biasing the mean and sd. so we need to set 99 to missing so the values go from 0 to 10 (see the slides on ‘preparing the data’ to do this).

Case study: descriptive statistics

Here some descriptive statistics for q25 where a value of ‘0’ or ’1’ represents ‘strongly oppose’ and value of ‘9’ or ’10’ represents ‘strongly favor’.

```
. tabstat   x25a x25b x25c x25d x25e x25f x25g x25h x25i x25j [aw=weight], s(mean median sd var count range min m
> ax)
```

stats	x25a	x25b	x25c	x25d	x25e	x25f	x25g	x25h	x25i	x25j
mean	5.909401	4.600287	5.747949	6.607931	6.193698	7.108907	6.302037	4.691252	5.790056	5.318128
p50	7	5	6	7	6	8	7	5	5	5
sd	4.078566	4.225217	3.623961	3.232957	3.011944	3.250066	3.129978	3.698315	3.007888	3.107059
variance	16.6347	17.85246	13.1331	10.45201	9.071804	10.56293	9.796765	13.67753	9.04739	9.653819
N	1042	1042	1034	1037	1013	1043	1038	1038	1027	1030
range	10	10	10	10	10	10	10	10	10	10
min	0	0	0	0	0	0	0	0	0	0
max	10	10	10	10	10	10	10	10	10	10

```
. tab elec gender [aw=weight], sum(x25c)

   Means, Standard Deviations, Frequencies and Number of Observations
of The government should provide health care coverage to all citizens who can't aff
```

RECODE of q5 (Q5. If the Presidential election were held today and the candidate	gender Male	gender Female	Total
Obama/Bid	8.1908614 2.0636006 192.45682 195	8.0364917 2.2601473 270.91296 281	8.1006079 2.1797639 463.36978 476
McCain/Pa	2.7627034 2.9422253 234.52383 254	3.4770192 3.230397 177.7414 203	3.070669 3.0867491 412.26523 457
Undecided	5.9889067 3.6637616 34.11981 40	5.8488136 3.2399565 53.313209 61	5.9034835 3.3944003 87.433019 101
Total	5.2670676 3.7346278 461.10046 489	6.1896809 3.4642965 501.96757 545	5.7479494 3.6239613 963.06803 1034

Here we use the combination tab/sum to explore a response to a third variable (usually continuous) in a crosstabulation. We are looking at the mean value of x25c (‘govt should provide health care’) by electoral preference and gender. For example, male Obama supporters tend to support government providing health care who can’t afford it (mean of 8.19). On the contrary, those who are male and prefer McCain tend to disagree (with a mean score of 2.76)

Case study: dummies

The quickest way to generate dummy variables is by using a combination of tab/gen command. Here is an example

```
. tab gender, gen(gender)

    gender |      Freq.     Percent        Cum.
-----------+-----------------------------------
      Male |        493       46.82       46.82
    Female |        560       53.18      100.00
-----------+-----------------------------------
     Total |      1,053      100.00

. tab1 gender*

-> tabulation of gender

    gender |      Freq.     Percent        Cum.
-----------+-----------------------------------
      Male |        493       46.82       46.82
    Female |        560       53.18      100.00
-----------+-----------------------------------
     Total |      1,053      100.00

-> tabulation of gender1

gender==Mal |
          e |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        560       53.18       53.18
          1 |        493       46.82      100.00
------------+-----------------------------------
      Total |      1,053      100.00

-> tabulation of gender2

gender==Fem |
        ale |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        493       46.82       46.82
          1 |        560       53.18      100.00
------------+-----------------------------------
      Total |      1,053      100.00
```

Case study: factor analysis

Factor analysis is a data reduction technique. Question 8 has a battery of questions evaluating favorability levels for different candidates/politicians

Variables

Principal-components factoring

```
. factor x8a x8b x8c x8d x8e x8f x8g x8h x8i x8j, pcf
(obs=897)

Factor analysis/correlation                       Number of obs    =        897
    Method: principal-component factors           Retained factors =          2
    Rotation: (unrotated)                         Number of params =         19
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.10910	1.89154	0.4109	0.4109
Factor2	2.21756	1.35886	0.2218	0.6327
Factor3	0.85870	0.12199	0.0859	0.7185
Factor4	0.73671	0.15331	0.0737	0.7922
Factor5	0.58340	0.05168	0.0583	0.8505
Factor6	0.53172	0.13910	0.0532	0.9037
Factor7	0.39262	0.11864	0.0393	0.9430
Factor8	0.27398	0.11808	0.0274	0.9704
Factor9	0.15591	0.01559	0.0156	0.9860
Factor10	0.14031	.	0.0140	1.0000

LR test: independent vs. saturated: chi2(45) = 4884.51 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
x8a	-0.9046	0.1045	0.1709
x8b	0.8586	0.2150	0.2165
x8c	-0.8531	0.1799	0.2399
x8d	0.9180	0.1434	0.1367
x8e	-0.4759	0.5533	0.4674
x8f	-0.1691	0.6717	0.5202
x8g	0.2197	0.5555	0.6432
x8h	0.8225	0.2936	0.2373
x8i	-0.0373	0.7252	0.4728
x8j	-0.0425	0.6554	0.5686

Total variance accounted by each factor. The sum of all eigenvalues = total number of variables.

When negative, the sum of eigenvalues = total number of factors (variables) with positive eigenvalues.

Kaiser criterion suggests to retain those factors with eigenvalues equal or higher than 1.

Since the sum of eigenvalues = total number of variables. Proportion indicate the relative weight of each factor in the total variance. For example, 4.109/10=0.4109. The first factor explains 41% of the total variance

Cumulative shows the amount of variance explained by n+(n-1) factors. For example, factor 1 and factor 2 account for 63% of the total variance.

Difference between one eigenvalue and the next.

Uniqueness is the variance that is 'unique' to the variable and not shared with other variables. It is equal to 1 – communality (variance that is shared with other variables). For example, 64% of the variance in 'x8g' is not share with other variables in the overall factor model. On the contrary 'x8a' has low variance not accounted by other variables (17%). Notice that the greater 'uniqueness' the lower the relevance of the variable in the factor model.

Factor loadings are the weights and correlations between each variable and the factor. The higher the load the more relevant in defining the factor's conceptual meaning. A negative value indicates an inverse impact on the factor. Here, two factors are retained because both have eigenvalues over 1. It seems that 'x8b', 'x8d' and 'x8h' define factor1, and 'x8f', and 'x8i' define factor2.

Case study: factor analysis

Factor analysis is a data reduction technique. Question 8 has a battery of questions evaluating favorability levels for different candidates/politicians

By default the rotation is varimax which produces orthogonal factors. This means that factors are not correlated to each other. This setting is recommended when you want to identify variables to create indexes or new variables without inter-correlated components

Same description as in the previous slide with new composition between the two factors. Still both factors explain 63% of the total variance observed.

The pattern matrix here offers a clearer picture of the relevance of each variable in the factor.

This is a correlation matrix between factor1 and factor2.

```
. rotate

Factor analysis/correlation                       Number of obs    =        897
    Method: principal-component factors           Retained factors =          2
    Rotation: orthogonal varimax (Kaiser off)     Number of params =         19
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	4.08288	1.83911	0.4083	0.4083
Factor2	2.24377	.	0.2244	0.6327

LR test: independent vs. saturated: chi2(**45**) = **4884.51** Prob>chi2 = **0.0000**

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
x8a	-0.8860	0.2103	0.1709
x8b	0.8780	0.1124	0.2165
x8c	-0.8260	0.2790	0.2399
x8d	0.9285	0.0343	0.1367
x8e	-0.4075	0.6055	0.4674
x8f	-0.0888	0.6869	0.5202
x8g	0.2836	0.5257	0.6432
x8h	0.8513	0.1947	0.2373
x8i	0.0483	0.7245	0.4728
x8j	0.0350	0.6559	0.5686

Factor rotation matrix

	Factor1	Factor2
Factor1	0.9930	-0.1177
Factor2	0.1177	0.9930

NOTE: If you want the factors to be correlated (oblique rotation) you need to use the option `promax` after `rotate`:
`rotate, promax`
Type `help rotate` for details.

Case study: factor analysis, step 3 (predict)

To create the new variables, after `factor`, `rotate` you type `predict`.

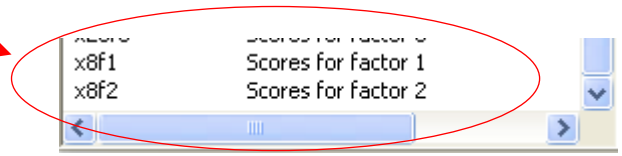
```
predict x8f1 x8f2 /*Or whatever name you prefer to identify the factors*/
```

```
. predict x8f1 x8f2
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2
x8a	-0.21306	0.07271
x8b	0.21892	0.07169
x8c	-0.19662	0.10498
x8d	0.22947	0.03792
x8e	-0.08565	0.26140
x8f	-0.00521	0.30564
x8g	0.08259	0.24245
x8h	0.21436	0.10790
x8i	0.02947	0.32580
x8j	0.02453	0.29473

These are the regression coefficients used to estimate the individual scores (per case/row)



We reduced all eight variables to two: x8f1 and x8f2. There is another way to use these results. We could create indexes out of each cluster of variables. For example, ‘x8b’, ‘x8d’ and ‘x8h’ define the first factor. You could aggregate these to create a new variable to measure ‘Republican favorability’. The second factor is defined by ‘x8e’, ‘x8f’, x8i’ and ‘x8j’ related to ‘government institutions’. Since all variables are in the same valence (go from 0 to 100), we can create the two new variables as

```
gen repubfav = (x8b + x8d + x8h)/3
gen govinst = (x8e + x8f + x8i + x8j)/4
```

Case study: regression

We use the command `regress` to run a regression

regress x8a gender age educ income x25*, robust

```
. regress x8a gender age educ income x25*, robust

Linear regression                               Number of obs =       857
                                                F( 14,   842) =    138.68
                                                Prob > F      =    0.0000
                                                R-squared     =    0.6114
                                                Root MSE      =     22.13

------------------------------------------------------------------------------
             |               Robust
         x8a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |    1.085681   1.524235     0.71   0.476    -1.906064    4.077427
         age |   -.0954027   .4441548    -0.21   0.830    -.9671832    .7763779
        educ |    1.570695   .8151773     1.93   0.054    -.0293229    3.170713
      income |   -.2996345   .4764621    -0.63   0.530    -1.234827    .6355583
        x25a |    1.101605   .2762611     3.99   0.000     .5593636    1.643846
        x25b |    .6041541   .2659564     2.27   0.023     .0821388    1.126169
        x25c |    2.749842   .3712377     7.41   0.000     2.021182    3.478502
        x25d |   -.1274084   .3054922    -0.42   0.677    -.7270241    .4722072
        x25e |   -.2741189   .2758408    -0.99   0.321    -.8155351    .2672973
        x25f |   -.9597492   .3174276    -3.02   0.003    -1.582792   -.3367069
        x25g |    1.201146   .3624039     3.31   0.001     .4898251    1.912467
        x25h |   -2.622509   .3181912    -8.24   0.000     -3.24705   -1.997968
        x25i |   -.6518584   .3177172    -2.05   0.041    -1.275469   -.0282476
        x25j |     .699863   .3073602     2.28   0.023     .0965809    1.303145
       x25f1 |  (dropped)
       x25f2 |  (dropped)
       x25f3 |  (dropped)
       _cons |    39.59818   7.345718     5.39   0.000     25.18011    54.01625
------------------------------------------------------------------------------
```

Case study: regression

Here is another example

regress x8b gender age educ income x25*, robust

```
. regress x8b gender age educ income x25*, robust

Linear regression                               Number of obs =       857
                                                F( 14,   842) =     70.66
                                                Prob > F      =    0.0000
                                                R-squared     =    0.4955
                                                Root MSE      =    22.135
```

x8b	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
gender	2.568956	1.529457	1.68	0.093	-.4330398	5.570951
age	-.3590177	.4220541	-0.85	0.395	-1.187419	.469384
educ	2.394501	.8376223	2.86	0.004	.7504277	4.038573
income	.7567806	.5004008	1.51	0.131	-.2253989	1.73896
x25a	-.4245393	.2513435	-1.69	0.092	-.9178727	.068794
x25b	-.5100364	.2616189	-1.95	0.052	-1.023538	.0034653
x25c	-1.546259	.3302899	-4.68	0.000	-2.194547	-.8979706
x25d	-.0041063	.2839938	-0.01	0.988	-.5615252	.5533125
x25e	-.5360159	.2764522	-1.94	0.053	-1.078632	.0066005
x25f	1.08052	.3298975	3.28	0.001	.4330022	1.728038
x25g	-.2805339	.3361083	-0.83	0.404	-.9402424	.3791746
x25h	3.539997	.3070789	11.53	0.000	2.937267	4.142727
x25i	.5077791	.3273211	1.55	0.121	-.134682	1.15024
x25j	-.0397483	.2948785	-0.13	0.893	-.6185315	.5390349
x25f1	(dropped)					
x25f2	(dropped)					
x25f3	(dropped)					
_cons	28.87047	7.224851	4.00	0.000	14.68964	43.0513

Case study: regression (exporting results)

	(1)	(2)
COEFFICIENT	x8a	x8b
gender	1.09	2.57
	(1.52)	(1.53)
age	-0.10	-0.36
	(0.44)	(0.42)
educ	1.57	2.39**
	(0.82)	(0.84)
income	-0.30	0.76
	(0.48)	(0.50)
x25a	1.10***	-0.42
	(0.28)	(0.25)
x25b	0.60*	-0.51
	(0.27)	(0.26)
x25c	2.75***	-1.55***
	(0.37)	(0.33)
x25d	-0.13	-0.00
	(0.31)	(0.28)
x25e	-0.27	-0.54
	(0.28)	(0.28)
x25f	-0.96**	1.08**
	(0.32)	(0.33)
x25g	1.20***	-0.28
	(0.36)	(0.34)
x25h	-2.62***	3.54***
	(0.32)	(0.31)
x25i	-0.65*	0.51
	(0.32)	(0.33)
x25j	0.70*	-0.04
	(0.31)	(0.29)
Constant	39.60***	28.87***
	(7.35)	(7.22)
Observations	857	857
R-squared	0.61	0.50
Adj. R-squared	0.60	0.49

Robust standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Use the `outreg2` command to export the output in a journal-paper like presentation. Run `outreg2` after each regression as follows

```
regress x8a gender age educ income x25*, robust

outreg2 using model, bdec(2) tdec(2) rdec(2) adec(2)
alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a))
word
```

`"model.rtf"` ← Click here to see the document

```
regress x8b gender age educ income x25*, robust

outreg2 using model, bdec(2) tdec(2) rdec(2) adec(2)
alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a))
word append
```

`"model.rtf"` ← Click here to see the document

Case study: do-file (part1)

```
log using workshop.log

/*Distribution of electoral preferences (frequencies)*/

tab q5  /*No weights*/
tab q5 [aweight=weight] /*With weights*/

tab qa  /*No weights*/
tab qa [aweight=weight] /*With weights*/

/*Electoral preferences by some demographics*/

tab q5 qa [aw=weight], col row /*Electoral preferences by gender*/
tab q5 f1 [aw=weight], col row /*Electoral preferences by age*/
tab q5 f4 [aw=weight], col row /*Electoral preferences by education*/
tab q5 f13 [aw=weight], col row /*Electoral preferences by income*/
tab q5 f8 [aw=weight], col row /*Electoral preferences by employment status*/

/*Preparing age variable*/
gen age=f1
tab age
labelbook f1
label value age f1
tab age
replace age=. if age>8
label variable age "Age"
tab age

/*Preparing education variable*/

gen educ=f4
tab educ
labelbook f4
label value educ f4
tab educ
replace educ=. if educ==8
label variable educ "Educational attainment"
tab educ
```

```
/*Preparing income variable*/
gen income=f13
tab income
labelbook f13
label value income f13
tab income
replace income=. if income==8
label variable income "Family income"
tab income

/*Preparing employment variable*/
gen employ=f8
tab employ
labelbook f8
label value employ f8
tab employ
replace employ=. if employ==8
label variable employ "Employment status"
tab employ

/*Preparing gender variable*/
gen gender=qa
tab gender
labelbook qa
label value gender qa
tab gender

/*Recoding electoral question*/
recode q5 (1=1 "Obama/Biden") (2=2 "McCain/Palin") (3 4 8=3
"Undecided/DK/NA/Other"), gen(elec) label(elec)
tab q5
tab elec

/*Testing for associations*/
tab elec gender, nofreq chi2
tab elec age, nofreq chi2
tab elec educ, nofreq chi2
tab elec income, nofreq chi2
tab elec employ, nofreq chi2
```

Case study: do-file (part 2)

```
/*Factor, data preparation*/

gen x8a = q8a
gen x8b = q8b
gen x8c = q8c
gen x8d = q8d
gen x8e = q8e
gen x8f = q8f
gen x8g = q8g
gen x8h = q8h
gen x8i = q8i
gen x8j = q8j

replace x8a = . if x8a>100
replace x8b = . if x8b>100
replace x8c = . if x8c>100
replace x8d = . if x8d>100
replace x8e = . if x8e>100
replace x8f = . if x8f>100
replace x8g = . if x8g>100
replace x8h = . if x8h>100
replace x8i = . if x8i>100
replace x8j = . if x8j>100

label variable x8a "Obama"
label variable x8b "McCain"
label variable x8c "Biden"
label variable x8d "Palin"
label variable x8e "Congress"
label variable x8f "Congressman"
label variable x8g "Supreme court"
label variable x8h "Pres. Bush"
label variable x8i "State gov"
label variable x8j "Local gov"

/*Running factor analysis */

factor  x8a x8b x8c x8d x8e x8f x8g x8h x8i x8j, pcf
rotate
predict x8f1 x8f2

gen repubfav = (x8b + x8d + x8h)/3
gen govinst = (x8e + x8f + x8i + x8j)/4
```

```
/*Descriptive statistics*/

tabstat  q8a x8a q8b x8b, s(mean)

tabstat  x8a x8b x8c x8d x8e x8f x8g x8h x8i x8j, s(mean median sd var count
range min max)

describe x8*

/* One more factor example */

gen x25a = q25a
gen x25b = q25b
gen x25c = q25c
gen x25d = q25d
gen x25e = q25e
gen x25f = q25f
gen x25g = q25g
gen x25h = q25h
gen x25i = q25i
gen x25j = q25j

replace x25a = . if x25a>10
replace x25b = . if x25b>10
replace x25c = . if x25c>10
replace x25d = . if x25d>10
replace x25e = . if x25e>10
replace x25f = . if x25f>10
replace x25g = . if x25g>10
replace x25h = . if x25h>10
replace x25i = . if x25i>10
replace x25j = . if x25j>10
```

Case study: do-file (part 3)

```
label variable x25a "A woman should be able to get an abortion if she wants one in the first three months of pregnancy, no matter what the reason"
label variable x25b "Gay couples should be allowed to marry, giving them full legal rights of married couples"
label variable x25c "The government should provide health care coverage to all citizens who can't afford it, even if it means higher taxes"
label variable x25d "Government regulation of financial institutions should be greatly increased"
label variable x25e "The government should have let financial institutions that got into trouble over bad mortgage debt go out of business rather than trying to rescue them"
label variable x25f "The government should allow offshore drilling for oil and gas in the waters off the U.S. coast        "
label variable x25g "Congress should pass stricter laws to protect the environment and reduce global warming, even if the economic costs are high"
label variable x25h "Our troops should stay in Iraq without a timetable for withdrawal until the Iraqi government is stable"
label variable x25i "Government should cut taxes on businesses to help the economy"
label variable x25j "The government should help people who can't afford their mortgage payments by suspending foreclosures until the economy has improved"

factor x25a x25b x25c x25d x25e x25f x25g x25h x25i x25j, pcf
rotate
predict x25f1 x25f2 x25f3


/*Regression*/

regress x8a gender age educ income x25*, robust
regress x8b gender age educ income x25*, robust
```

Exploring data: annotated output

Exploring data: frequencies (intro)

Frequency refers to the number of times a value is repeated. Frequencies are usually used to analyze categorical data. The tables below are *frequency tables.* Values are in ascending order. Use the command `tab` (type `help tab` for more details)

```
. tab major
```

Major	Freq.	Percent	Cum.
Econ	10	33.33	33.33
Math	10	33.33	66.67
Politics	10	33.33	100.00
Total	30	100.00	

'Freq.' provides a raw count of each value. In this case 10 students for each major.

'Percent' gives the relative frequency for each value. For example, 33.33% of the students in this group are econ majors.

'Cum.' is the cumulative frequency in ascending order of the values. For example, 66.67% of the students are econ or math majors.

```
. tab  readnews
```

Newspaper read / week	Freq.	Percent	Cum.
3	6	20.00	20.00
4	5	16.67	36.67
5	9	30.00	66.67
6	7	23.33	90.00
7	3	10.00	100.00
Total	30	100.00	

'Freq.' Here 6 students read the newspaper 3 days a week, 9 students read it 5 days a week.

'Percent'. Those who read the newspaper 3 days a week represent 20% of the sample, 30% of the students in the sample read the newspaper 5 days a week.

'Cum.' 66.67% of the students read the newspaper 3 to 5 days a week.

Exploring data: crosstabs

Also known as *contingency tables*, crosstabs help you to analyze the relationship between two or more variables (mostly categorical). Below is a crosstab between the variable 'ecostatu' and 'gender'. We use the command `tab` (with two variables to make the crosstab).

Options 'col', 'row' gives you the column and row percentages.

var1 var2

```
. tab ecostatu gender, col row
```

Key

frequency
row percentage
column percentage

Status of Nat'l Eco	Gender of Respondent male	female	Total
Very well	90 60.40 14.33	59 39.60 7.92	149 100.00 10.85
Fairly well	337 50.30 53.66	333 49.70 44.70	670 100.00 48.80
Fairly badly	139 39.94 22.13	209 60.06 28.05	348 100.00 25.35
Very badly	57 29.84 9.08	134 70.16 17.99	191 100.00 13.91
Not sure	2 16.67 0.32	10 83.33 1.34	12 100.00 0.87
Refused	3 100.00 0.48	0 0.00 0.00	3 100.00 0.22
Total	628 45.74 100.00	745 54.26 100.00	1,373 100.00 100.00

The first value in a cell tells you the number of observations for each xtab. In this case, 90 respondents are 'male' and said that the economy is doing 'very well', 59 are 'female' and believe the economy is doing 'very well'

The second value in a cell gives you row percentages for the first variable in the xtab. Out of those who think the economy is doing 'very well', 60.40% are males and 39.60% are females.

The third value in a cell gives you column percentages for the second variable in the xtab. Among males, 14.33% think the economy is doing 'very well' while 7.92% of females have the same opinion.

You can use `tab1` for multiple frequencies or `tab2` to run all possible crosstabs combinations. Type `help tab` for further details.

Exploring data: crosstabs (a closer look)

You can use crosstabs to compare responses among categories in relation to aggregate responses. In the table below we compare male and female responses vs. the national aggregate.

```
. tab ecostatu gender, col row

Key

frequency
row percentage
column percentage
```

Status of Nat'l Eco	Gender of Respondent male	female	Total
Very well	90 60.40 14.33	59 39.60 7.92	149 100.00 10.85
Fairly well	337 50.30 53.66	333 49.70 44.70	670 100.00 48.80
Fairly badly	139 39.94 22.13	209 60.06 28.05	348 100.00 25.35
Very badly	57 29.84 9.08	134 70.16 17.99	191 100.00 13.91
Not sure	2 16.67 0.32	10 83.33 1.34	12 100.00 0.87
Refused	3 100.00 0.48	0 0.00 0.00	3 100.00 0.22
Total	628 45.74 100.00	745 54.26 100.00	1,373 100.00 100.00

As a rule-of-thumb, a margin of error of ±4 percentage points can be used to indicate a significant difference (some use ±3).

For example, rounding up the percentages, 11% (10.85) answer 'very well' at the national level. With the margin of error, this gives a range roughly between 7% and 15%, anything beyond this range could be considered significantly different (remember this is just an approximation). It does not appear to be a significant bias between males and females for this answer.

In the 'fairly well' category we have 49%, with range between 45% and 53%. The response for males is 54% and for females 45%. We could say here that males tend to be a bit more optimistic on the economy and females tend to be a bit less optimistic.

If we aggregate responses, we could get a better picture. In the table below 68% of males believe the economy is doing well (comparing to 60% at the national level, while 46% of females thing the economy is bad (comparing to 39% aggregate). Males seem to be more optimistic than females.

RECODE of ecostatu (Status of Nat'l Eco)	Gender of Respondent male	female	Total
Well	427 52.14 67.99	392 47.86 52.62	819 100.00 59.65
Bad	196 36.36 31.21	343 63.64 46.04	539 100.00 39.26
Not sure/ref	5 33.33 0.80	10 66.67 1.34	15 100.00 1.09
Total	628 45.74 100.00	745 54.26 100.00	1,373 100.00 100.00

```
recode ecostatu (1 2 = 1 "Well") (3 4 = 2 "Bad") (5 6=3 "Not sure/ref"), gen(ecostatu1) label(eco)
```

Exploring data: crosstabs (test for associations)

To see whether there is a relationship between two variables you can choose a number of tests. Some apply to nominal variables some others to ordinal. I am running all of them here for presentation purposes.

Likelihood-ratio χ^2(chi-square) | *Goodman & Kruskal's γ (gamma)* | X^2(chi-square) | Cramer's *V* | Kendall's τ_b (tau-b)

```
. tab ecostatu1 gender, col row nokey chi2 lrchi2 V exact gamma taub

Enumerating sample-space combinations:
stage 3:  enumerations = 1
stage 2:  enumerations = 16
stage 1:  enumerations = 0
```

Fisher's exact test

RECODE of ecostatu (Status of Nat'l Eco)	Gender of Respondent male	female	Total
well	427	392	819
	52.14	47.86	100.00
	67.99	52.62	59.65
Bad	196	343	539
	36.36	63.64	100.00
	31.21	46.04	39.26
Not sure/ref	5	10	15
	33.33	66.67	100.00
	0.80	1.34	1.09
Total	628	745	1,373
	45.74	54.26	100.00
	100.00	100.00	100.00

Pearson chi2(2) =	33.5266	Pr = 0.000
likelihood-ratio chi2(2) =	33.8162	Pr = 0.000
Cramér's V =	0.1563	
gamma =	0.3095	ASE = 0.050
Kendall's tau-b =	0.1553	ASE = 0.026
Fisher's exact =		0.000

- For *nominal* data use chi2, lrchi2, V
- For *ordinal* data use gamma and taub
- Use `exact` instead of `chi2` when frequencies are less than 5 across the table.

X^2(chi-square) tests for relationships between variables. The null hypothesis (Ho) is that there is no relationship. To reject this we need a Pr < 0.05 (at 95% confidence). Here both chi2 are significant. Therefore we conclude that there is some relationship between perceptions of the economy and gender

Cramer's *V* is a measure of association between two nominal variables. It goes from 0 to 1 where 1 indicates strong association (for rXc tables). In 2x2 tables, the range is -1 to 1. Here the V is 0.15, which shows a small association.

Gamma and taub are measures of association between two ordinal variables (both have to be in the same direction, i.e. negative to positive, low to high). Both go from -1 to 1. Negative shows inverse relationship, closer to 1 a strong relationship. Gamma is recommended when there are lots of ties in the data. Taub is recommended for square tables.

Fisher's exact test is used when there are very few cases in the cells (usually less than 5). It tests the relationship between two variables. The null is that variables are independent. Here we reject the null and conclude that there is some kind of relationship between variables

Exploring data: descriptive statistics

For continuous data we use descriptive statistics. These statistics are a collection of measurements of two things: *location* and *variability*. Location tells you the central value of your variables (the mean is the most common measure of this) . 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 such variability. We use the command `tabstat` to get these stats (the 's' after the comma means 'statistics').

```
. tabstat    age sat score height readnews, s(mean median sd var count range min max )
```

stats	age	sat	score	height	readnews
mean	25.2	1848.9	80.36667	66.43333	4.866667
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
N	30	30	30	30	30
range	21	971	33	16	4
min	18	1338	63	59	3
max	39	2309	96	75	7

- The *mean* is the sum of the observations divided by the total number of observations.
- The *median* (p50 in the table above) is the number in the middle . To get the median you have to order the data from lowest to highest. 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.
- 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 from the mean, 95% within 2 sd and 99% within 3 sd
- The *variance* measures the dispersion of the data from the mean. It is the simple mean of the squared distance from the mean.
- *Count* (N in the table) refers to the number of observations per variable.
- *Range* is a measure of dispersion. It is the difference between the largest and smallest value, max – min.
- *Min* is the lowest value in the variable.
- *Max* is the largest value in the variable.

Exploring data: regression (what to look for)

Lets run the regression:

```
regress x8a gender age educ income, robust
```

Dependent variable (Y)

Independent variables (X)

To control for heteroskedasticity

```
. regress x8a gender age educ income, robust

Linear regression                               Number of obs =     923
                                                F(  4,   918) =   15.43
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0593
                                                Root MSE      =  34.364

------------------------------------------------------------------------------
             |               Robust
         x8a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   9.492816   2.273604     4.18   0.000      5.03075    13.95488
         age |   -.846633   .6476339    -1.31   0.191    -2.117648    .4243819
        educ |   6.994452   1.099606     6.36   0.000     4.836419    9.152486
      income |  -2.325835   .7030282    -3.31   0.001    -3.705564   -.946106
       _cons |   25.88266   7.899246     3.28   0.001     10.37998    41.38533
------------------------------------------------------------------------------
```

This is the p-value of the model. It indicates the reliability of X to predict Y. Usually we need a p-value lower than 0.05 to show a statistically significant relationship between X and Y.

R-square shows the amount of variance of Y explained by X. In this case the model explains 5% of the variance in x8a.

x8a = 25.88 + 9.5*gender – 0.8*age
+ 6.9*educ -2.3*income

The t-values test the hypothesis that the coefficient is different from 0. To reject this, you need a t-value greater than 1.96 (95% confidence). You can get the t-values by dividing the coefficient by its standard error. The t-values also show the importance of a variable in the model. In this case, educ is the most important.

These are two-tail p-values for each coefficient. It tests the hypothesis that the coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (you could choose also an alpha of 0.01). In this case, only “age” does not seem to be significant.

Exploring data: regression, publishing regression output (outreg2)

Once you define your final model, you can export your regression results using either your log file or the option outreg2. For the log you just open it using any word processor and copy-and-paste the regression table into excel or word. The command outreg2 gives you the type of presentation you see in scholar's papers. Let's say the final regression is

```
regress csat percent percent2 high
```

After running the regression type the following if you want to export the results to excel*

```
outreg2 using results, bdec(2) tdec(2) rdec(2) adec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a)) excel
```

Or this if you want to export to word

```
outreg2 using results, bdec(2) tdec(2) rdec(2) adec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a)) word
```

You will see this in Stata's output window

For excel

```
. outreg2 using results, bdec(2) tdec(2) rdec(2) adec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e
> (r2_a)) excel
"results.xml"
seeout
```

Click here to see the output, a excel/word window will open

For word

```
. outreg2 using results, bdec(2) tdec(2) rdec(2) adec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r
> 2_a)) word
"results.rtf"
seeout
```

Click on `seeout` to browse the results

Name of the file for the output

Set # of decimals for coefficients

Set # of decimals for auxiliary statistics

Set # of decimals for the R^2

Set # of decimals for added statistics (`addstat` option)

Levels of significance

Include some additional statistic, in this case adj. R-sqr. You can select any statistics on the return lists (e-class, r-class or s-class). After running the regression type `ereturn list` for a list of available statistics.

Type `help outreg2` for more details. If you do not see `outreg2`, you may have to install it by typing `ssc install outreg2`. If this does not work type `findit outreg2`, select from the list and click "install".

Note: If you get the following error message (when you use the option `append` or `replace` it means that you need to close the excel/word window.

```
file results.rtf is read-only; cannot be modified or erased
```

*See the following document for some additional info/tips http://www.fiu.edu/~tardanic/brianne.pdf

Exploring data: regression, publishing regression output (`outreg2`)

This is how the output would like (you will still need to do some additional editing):

In excel

	A	B
1	v1	v2
2	COEFFICIENT	csat
3		
4	percent	-6.52***
5		(0.51)
6	percent2	0.05***
7		(0.01)
8	high	2.99***
9		(0.49)
10	Constant	844.82***
11		(36.63)
12	Observations	51
13	R-squared	0.93
14	Adj. R-squared	0.92
15	Standard errors in parentheses	
16	*** p<0.001, ** p<0.01, * p<0.05	

In word

COEFFICIENT	csat
percent	-6.52***
	(0.51)
percent2	0.05***
	(0.01)
high	2.99***
	(0.49)
Constant	844.82***
	(36.63)
Observations	51
R-squared	0.93
Adj. R-squared	0.92

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

You can add more models to compare. Lets say you want to add another model without `percent2`:

```
regress csat percent high
```

Now type to export the results to excel (**notice** we add the `append` option)

```
outreg2 using results, bdec(2) tdec(2) rdec(2) adec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a)) excel append
```

In excel

	A	B	C
1	v1	v2	v3
2		(1)	(2)
3	COEFFICIENT	csat	csat
4			
5	percent	-6.52***	-2.32***
6		(0.51)	(0.16)
7	percent2	0.05***	
8		(0.01)	
9	high	2.99***	2.56**
10		(0.49)	(0.76)
11	Constant	844.82***	831.63***
12		(36.63)	(57.39)
13	Observations	51	51
14	R-squared	0.93	0.81
15	Adj. R-squared	0.92	0.80
16	Standard errors in parentheses		
17	*** p<0.001, ** p<0.01, * p<0.05		

In word

	(1)	(2)
COEFFICIENT	csat	csat
percent	-6.52***	-2.32***
	(0.51)	(0.16)
percent2	0.05***	
	(0.01)	
high	2.99***	2.56**
	(0.49)	(0.76)
Constant	844.82***	831.63***
	(36.63)	(57.39)
Observations	51	51
R-squared	0.93	0.81
Adj. R-squared	0.92	0.80

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05