Stata0, Introduction

Yahya Salimi
Ph.D. in Epidemiology,
Department of Epidemiology,
Kermanshah University of Medical Science

What is STATA?

- A general purpose statistical analysis package used by
  - epidemiologists, demographers, clinical researchers, social scientists, many others
- Tool to graphically display data
  - Good for data exploration
  - Also good for publishing in journals

Why STATA?

- Easy to learn
- Powerful
- It will help you produce papers

Anatomy of A Clinical Research Project

- Collect (the data)
- Clean
- Explore
- Analyze
- Submit (for publication)
- Revise

Collect the Data

- STATA is good for analyzing
  - large secondary databases
  - smaller home grown data
- Store the data as a relational database (or maybe as a spreadsheet)
  - It's easy to convert to STATA format from SAS and other formats

Clean the Data

- Merge in other sources of data
  - STATA does merges of all types, including match merge, table-lookup, and more complicated merging
- Recode variables
- Hunt for outliers
- Apply inclusion/exclusion criteria
- Treat missing variables consistently

Explore the Data

- Make a data codebook
- Examine univariate statistics
  - mean, standard deviation, percentiles
- Explore bivariate relationships
  - correlations, conditional means, etc.
- Examine the data graphically
  - STATA has powerful graphics capabilities (with a simple interface)

Analyze the Data

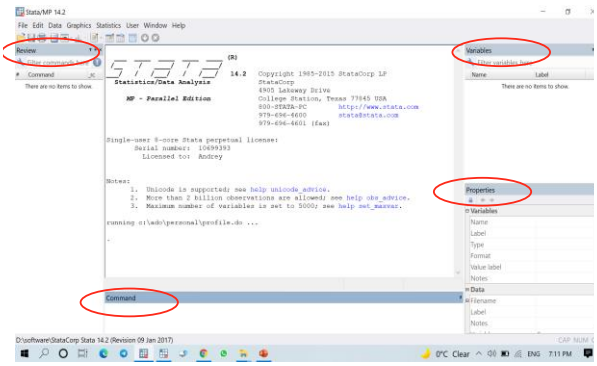
- STATA is powerful all-purpose statistical package with most common statistical computations built in
- STATA is extensible for uncommon statistical computations
  - You can share the tools you develop with the rest of the STATA community
  - Built-in and user written commands have a common interface
  - The STATA community is vibrant and helpful

Data handling

Import data

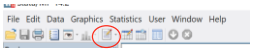
- Using SPSS
  - Save as, Stata Version 8 SE
- Using Excel
- Text Data
  - Type your own data

Interface

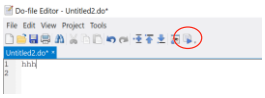


Do Editor

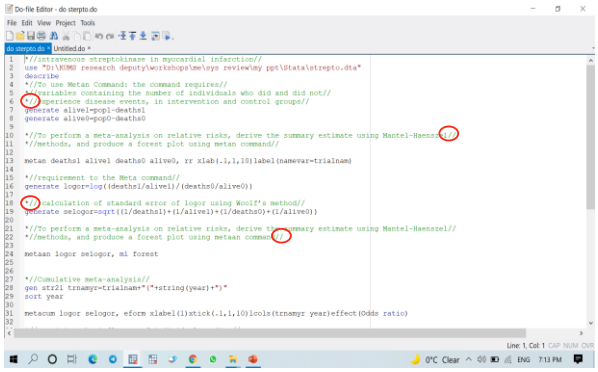
- New
  - Ctrl-9, or:



- Run
  - Mark commands, Ctrl-D to do (execute)



Do-file example



Syntax

- Syntax
  [bysort varlist:] command [varlist] [if exp] [in range][, opts]
- Examples
  - mean age
  - mean age if sex==1
  - bysort sex: summarize age
  - summarize age ,detail

DATA MANAGEMENT

- If you are only interested in a subset of your data, you can inspect it using filters. E.g. If you are only interested in price of a particular type of car you can type:
  - sum if price>=3000 & price<=4400
  - sum if mpg>=16& mpg<=23
- And then you can contrast
  - sum if price>=3000 |price<=4400
  - sum if mpg>=16 |mpg<=23
- Interpretation of Logical Operators in STATA.

>=	greater or equal to
<=	less or equal to
==	equal to
&	and
|	or
!= or ~=	not equal to
>	greater than
<	Less than
.	missing

Use and save data

- Open data
  - set memory 200m
  - use “C:\Course\Myfile.dta”, clear
- Describe
  - describe describe all variables
  - list x1 x2 in 1/20 list obs nr 1 to 20
- Save data
  - save “C:\Course\Myfile.dta” ,replace

Drop and keep

- Drop
  - drop x1 x2 — drop variables x1 and x2
  - drop if sex==1 — drop males
  - drop if age==. — drop missing
- Keep
  - same as drop

Recode

- Syntax
  - From 4 to 2 groups:
    recode educ (1 2=1) (3 4=2)(missing=.), gen(educ2)
  - **From cont. to 3 groups:**
    recode age (min/19=1) (20/29=2) (30/max=3), gen(age3)

Labels

- Variable
  - label variable q1 ”Age”
- Value
  1) label define freqLab 1”Low” 2”Med” 3”High”
  2a) label values smoke freqLab
  2b) label values drink freqLab
- List
  - label list

Generate, replace

- Age square
  - generate ageSqr=age^2
- Young/Old
  - generate old=0 if (age<=50)
  - replace old=1 if (age>50)
- Observation numbers
  - gen id=_n
  - gen lag=age[_n-1]

Missing

- Obs!!!
  - Missing values are large numbers
  - age>30 will include missing.
  - age>30 if age<. will not.
- Test
  - replace x=0 if (x==.)
- Remove
  - drop if age==.
- Change
  - replace educ=. if educ==99

Describe missing

- Summarize variables

```
summarize id bullied sex
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	2079	5136.855	2978.587	1	10308
bullied	2011	.1700646	.375783	0	1
sex	2050	1.488293	.4999849	1	2

- Missing in tables

```
. tab bullied sex, missing                    /* 2-way
```

Is bullied	Child's sex: Boy	Girl	.	Total
no	806	839	24	1,669
yes	208	129	5	342
.	35	33	0	68
Total	1,049	1,001	29	2,079

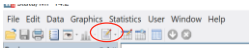
Handle data with many variables

- Describe
  - describe vars — format and labels
  - summarize vars — N, mean, std, min and max
  - codebook vars — range, missing, mean and std, percentiles
- Find variables
  - describe, simple — list all variables
  - lookfor age — list variables with “age” in name or label
  - describe age*, n — list vars starting with “age” and show var number
- Change order
  - order vars — change order of variables

Help

- General
  - help *command*
  - findit *keyword* — *search Stata+net*
- Examples
  - help table
  - findit aflogit

Summing up



- Use do files
  - Mark, Ctrl-D to do (execute)
- Syntax
  - command [varlist] [if exp] [in range] [, options]
- Missing
  - age>30 & age<.
  - generate old=(age>50) if age<.
- Help
  - help describe

Books

Data Analysis Using Stata
by Ulrich Kohler and Frauke Kreuter

Statistics with Stata (Updated for Version 9)
by Lawrence C. Hamilton

A visual guide to Stata graphics
by M.N. Mitchell

Multilevel and longitudinal modeling using Stata
by S. Rabe-Hesketh, A. Skrondal

Stata 1, Graphics

Yahya Salimi
Ph.D. in Epidemiology,
Department of Epidemiology,
Kermanshah University of Medical Science

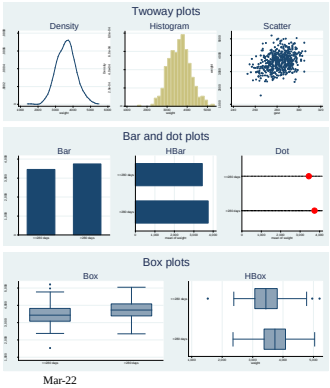
Why use graphs?

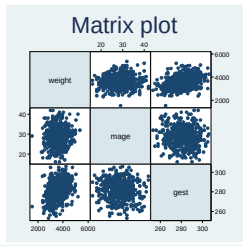
Structure of talk

- Order
  - Work/presentation plots
  - Plot types
  - Outcome type
- Focus:
  - The right plot
  - The commands

Plot types

Plottypes



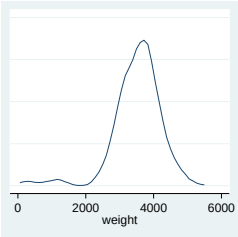


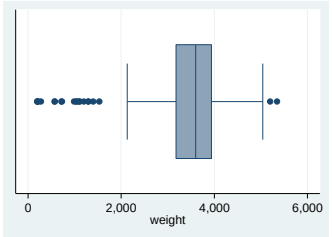
Mar-22 7

Continuous outcome

Univariate

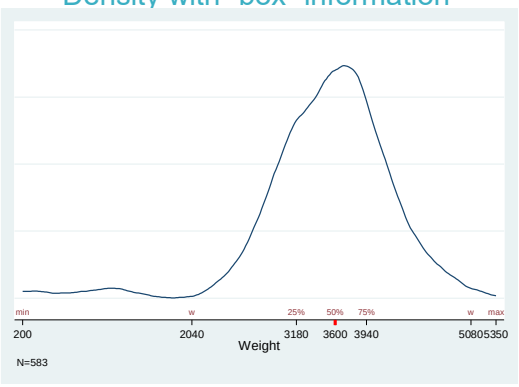
- Density
  - kdensity weight
- Boxplot
  - graph hbox weight





Mar-22 9

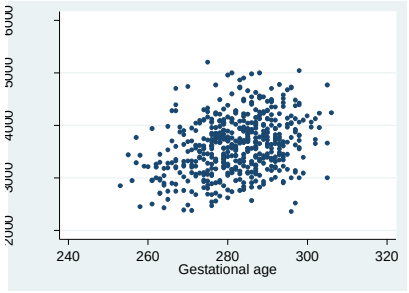
Density with “box” information



Mar-22 10

Bivariate

- Scatter
  - scatter weight gest



Twoway density

- Syntax
  - graph twoway (plot1, opts) (plot2, opts), opts
- One plot
  - kdensity x
- Two plots overlaid
  - twoway (kdensity weight if sex==1, lcolor(blue)) ///
    (kdensity weight if sex==2, lcolor(red))
- Side by side
  - twoway (kdensity weight), by(sex)

Twoway scatter + fit

- Syntax
  - graph twoway (plot1, opts) (plot2, opts), opts
- Examples
  - scatter y x
  - twoway (scatter y x) (fpfitci y x) (lfit y x)

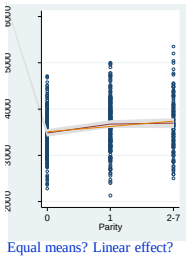
Fitlines

lfit	lfitci	Linear
qfit	qfitci	quadratic
mband, mspline		Median band, median spline
	fpfitci	Fractional polynomial
lowess		Local regression

Continuous by 3 categories

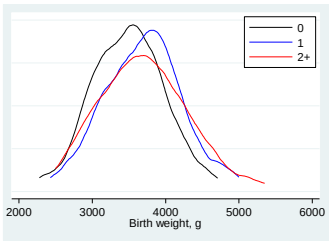
- Is birth weight the same over parity?

Scatterplot



Equal means? Linear effect?
Outliers?

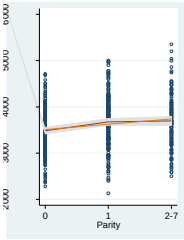
Density plot



Equal variances?

Continuous by 3 categories

Scatterplot



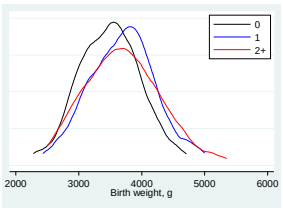

```
twoway (scatter weight parity3)
       (fpfitci  weight parity3)
       (lfit     weight parity3)
       , legend(off)
```

- Look for:
  - Outliers (all analyses)
  - Non-linear effects (regression)

Continuous by 3 categories

Density plot

```
twoway
(kdensity weight if parity3==0, lcol(black))
(kdensity weight if parity3==1, lcol(blue))
(kdensity weight if parity3==2, lcol(red))
, yscale(off)
```



- Look for:
  - Different locations
  - Different shapes (ANOVA, regression)

Twoway options

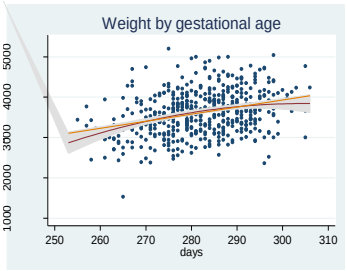
- Syntax
  - graph twoway (plot1, opts) (plot2, opts), opts
- Options

– lcolor(red)	line color
– lpattern(“.-”)	line pattern
– lwidth(*2)	line width *2
– legend(	
ring(0)	legend inside plot
pos(2)	legend at 2 o’clock position
col(1)	legends in 1 column
label(1 “First”)	legend label plot 1
label(2 “Second”)	legend label plot 2
)	

Continuous by continuous

- twoway
  (scatter weight gest)
  (fpfitci weight gest)
  (lfit weight gest)

- Look for:
  - Main effect (line)
  - Non-linearity (smooth)
  - outliers



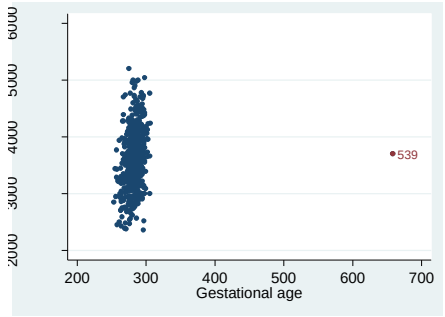
More twoway options

- Syntax
  - graph twoway (plot1, opts) (plot2, opts), opts
- Options
  - msize(*0.5) marker size
  - mlabel(id) marker label =variable id

  - xline(24) line at x=24
  - scale(1.5) all elements 1.5*larger

Mark outliers

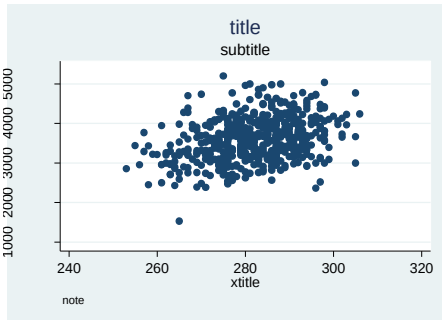
```
twoway (scatter weight gest)
       (scatter weight gest if gest>300, mlabel(id))
```



Titles, legend, labels and scale

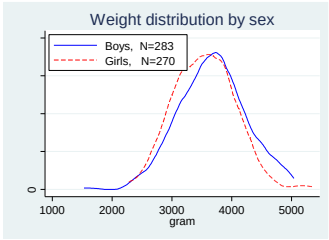
Titles

```
scatter weight gest, title("title") subtitle("subtitle")          ///
                     xtitle("xtitle") ytitle("ytitle") note("note")
```



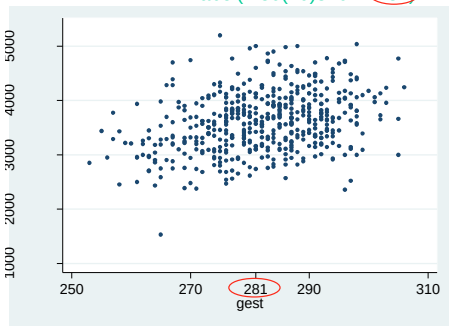
Legend

- …, legend(
  ring(0) pos(11) col(1)
  label(1 "Boys, N=283")
  label(2 "Girls, N=270"))
- …, legend(off)



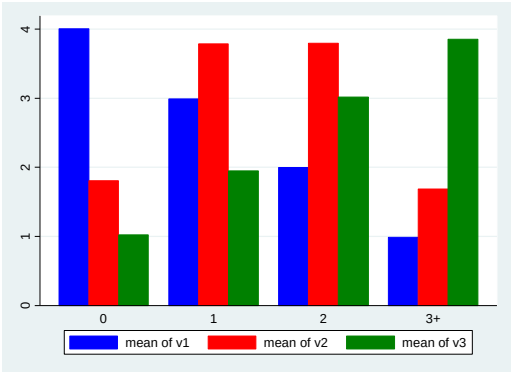
Axis scale and label

```
scatter weight gest, xscale(range(250 310))  ///
                     xlabel( 250(20)310  281)
```



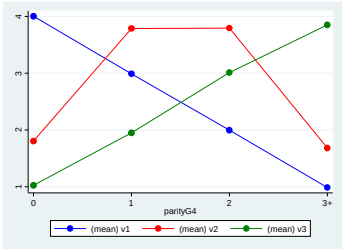
Categorical outcome

Comparing means or proportions



Comparing means/prop. better

```
preserve                                        "save" data
collapse  (mean)  v1 v2 v3, by(parity)          aggregate
list                                            list the new data
twoway (scatter v1 parity) (line v1 parity) ///
       (scatter v2 parity) (line v2 parity)  ///
       (scatter v3 parity) (line v3 parity)
restore                                         restore original data
```



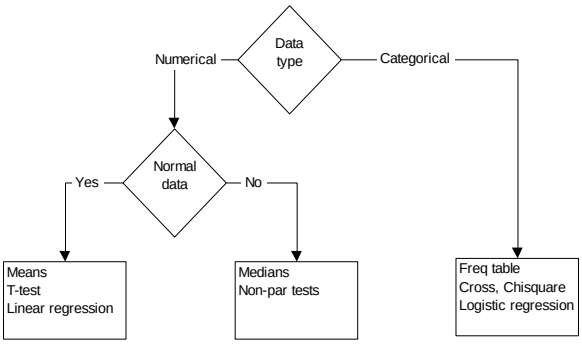
Stata 2, Bivariate analysis

Yahya Salimi
Ph.D. in Epidemiology,
Department of Epidemiology,
Kermanshah University of Medical Science

Datatypes

- Categorical data
  - Nominal: *married/ single/ divorced*
  - Ordinal: *small/ medium/ large*
- Numerical data
  - Discrete: *number of children*
  - Continuous: *weight*

Data type dictates type of analysis



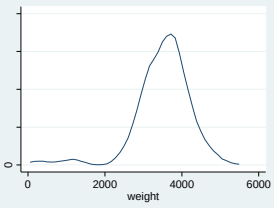
Continuous symmetric outcome

Example:

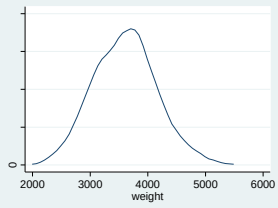
Birth weight

Distribution

kdensity weight



drop if weight<2000
kdensity weight



Central tendency and dispersion

Mean and standard deviation:

```
. summarize weight
```

Variable	Obs	Mean	Std. Dev.	Min	Max
weight	564	3603.883	543.5944	2130	5350

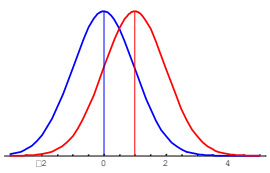
Mean with confidence interval:

```
. mean weight
Mean estimation                 Number of obs    =       564
```

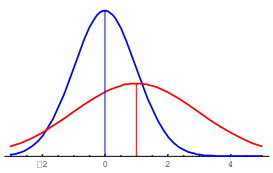
	Mean	Std. Err.	[95% Conf.	Interval]
weight	3603.883	22.88945	3558.924	3648.842

Compare groups, equal variance?

- Equal



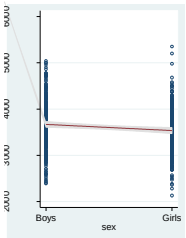
- Not equal



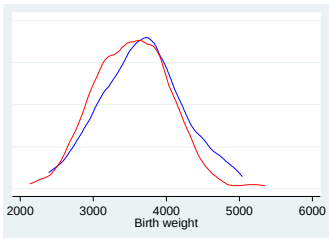
2 independent samples

Are birth weights the same for boys and girls?

Scatterplot



Density plot



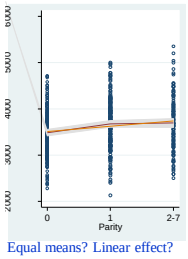
2 independent samples test

```
. ttest weight, by(sex)                                  /* T-test */
Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Boy |     291    3668.419    32.84396     560.276    3603.776    3733.062
    Girl |     273    3535.092    31.31681     517.4386   3473.437    3596.746
---------+--------------------------------------------------------------------
combined |     564    3603.883    22.88945     543.5944   3558.924    3648.842
---------+--------------------------------------------------------------------
    diff |            133.3277    45.49667                43.96337    222.692
------------------------------------------------------------------------------
    diff = mean(Boy) - mean(Girl)                                 t =   2.9305
Ho: diff = 0                                     degrees of freedom =      562

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9982         Pr(|T| > |t|) = 0.0035          Pr(T > t) = 0.0018
```

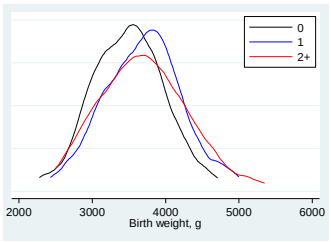

K independent samples

- Is birth weight the same over parity?

Scatterplot



Density plot



Equal means? Linear effect?
Outliers?

Equal variances?

K independent samples test

```
. oneway weight parity3, tabulate
   RECODE of |      Summary of weight
      parity |       Mean   Std. Dev.       Freq.
-------------+------------------------------------
           0 |       3485        491          225
           1 |       3677        544          215
         2-7 |       3695        598          123
-------------+------------------------------------
       Total |       3604        544          563

                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups    5334695.24         2   2667347.62      9.28     0.0001
 Within groups    160987259        560   287477.248
------------------------------------------------------------------------
       Total      166321954        562   295946.538

Bartlett's test for equal variances:  chi2(2) =   6.4740  Prob>chi2 =   0.039
```

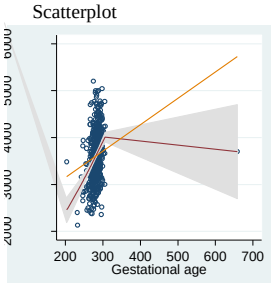
equal means?

Equal variances?

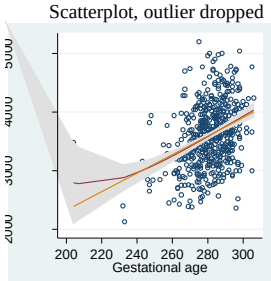
Continuous by continuous

- Does birth weight depend on gestational age?

Scatterplot



Scatterplot, outlier dropped



Continuous by continuous tests

- Cut gestational age up in groups, then use T-test or ANOVA

or

- Use linear regression with 1 covariate

Test situations

- 2 independent samples
  - ttest weight, by(sex)
- K independent samples
  - oneway weight parity
- By continuous
  - regress weight gestAge
- 2 dependent samples (Paired)
  - ttest weight_last_year = weight_today

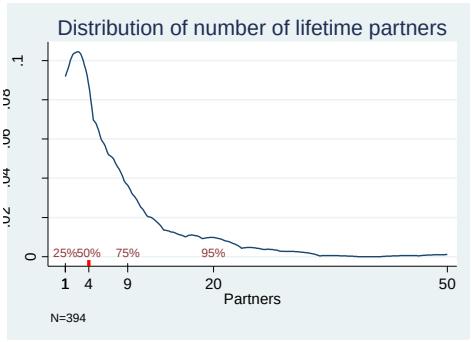
Continuous skewed outcome

Example:

Number of sexual partners

Distribution

kdensity partners if partners<=50



Central tendency and dispersion

Median and percentiles:

```
centile partners, centile(25 50 75) cci
```

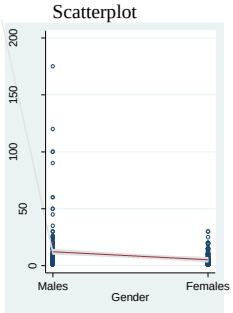
Variable	Obs	Percentile	Centile	Binomial Exact [95% Conf. Interval]	
partners	401	25	1	1	2
		50	4	4	5
		75	10	8	10

```
summarize partners, detail
```

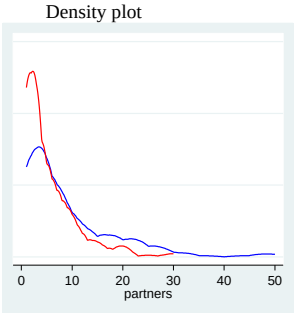

2 independent samples

Do males and females have the same number of partners?

Scatterplot



Density plot



2 independent samples test

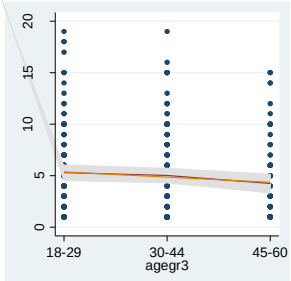
```
. tabstat partners, stat(p50) by(gender)
gender |       p50
-------+----------
  Male |         6
Female |         3
-------+----------
 Total |         4
-------------------

. ranksum partners, by(gender)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

      gender |       obs    rank sum    expected
-------------+---------------------------------
        Male |       179     40924.5     35979
      Female |       222     39676.5     44622
-------------+---------------------------------
    combined |       401       80601       80601

unadjusted variance  1331223.00
adjustment for ties    -28144.98
                     ----------
adjusted variance    1303078.02

Ho: partners(gender==Male) = partners(gender==Female)
                 z =   4.332
     Prob > |z|   =   0.0000
```

equal medians?

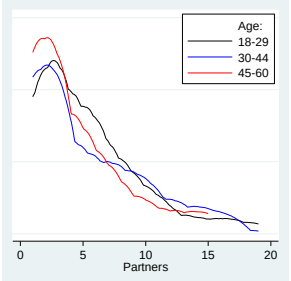
K independent samples

Do partners vary with age?

Scatterplot (partners<20)



Density plot (partners<20)



K independent samples test

```
. tabstat partners, stat(p50) by(agegr3)
agegr3 |       p50
-------+----------
18-29  |         5
30-44  |       4.5
45-60  |         3
-------+----------
Total  |         4
-------------------

. kwallis partners, by(agegr3)

Test: Equality of populations (Kruskal-Wallis test)

    agegr3 |  Obs | Rank Sum
-----------+------+----------
     18-29 |  140 | 29291.50
     30-44 |  160 | 31512.50
     45-60 |   94 | 17011.00
-----------+------+----------

chi-squared =     3.469 with 2 d.f.
probability =     0.1765

chi-squared with ties =     3.541 with 2 d.f.
probability =     0.1702
```

equal medians?

Table of tests

	Numerical data		Proportions
	Normal	**Skewed**	
1 sample	One sample T-test	Kolmogorov-Smirnov	Binomial
2 independent samples	Independent sample T-test	Mann-Whitney U	Chi-square
K independent samples	ANOVA	Kruskal-Wallis	Chi-square
2 dependent samples	Paired sample T-test	Wilcoxon signed rank test	Mc-Nemar (2x2)

Categorical ordered: use nonparametric tests

Categorical data

Example:

Being bullied

Frequency and proportion

Frequency:

```
. tabulate bullied
Is bullied |      Freq.     Percent        Cum.
-----------+-----------------------------------
         no |      1,669       82.99       82.99
        yes |        342       17.01      100.00
-----------+-----------------------------------
      Total |      2,011      100.00
```

Proportion with CI:

```
. proportion bullied
Proportion estimation                 Number of obs     =       2011

                                                Binomial Wald
             |   Proportion   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------
bullied      |
          no |   .8299354   .0083798       .8135014    .8463693
         yes |   .1700646   .0083798       .1536307    .1864986
```

Proportion, confidence interval

proportion: $p = \frac{x}{n}$

x="disease"
n=total number

standard error: $se(p) = \sqrt{\frac{p(1-p)}{n}}$

confidence interval: $CI(p) = p \pm 2se(p)$

Crosstables

Are boys bullied as much as girls?

```
. tabulate bullied sex,col chi2 nofreq
```

Is bullied	Child's sex: Boy	Girl	Total
no	79.49	86.67	83.00
yes	20.51	13.33	17.00
Total	100.00	100.00	100.00

Pearson chi2(1) = 18.1234 Pr = 0.000

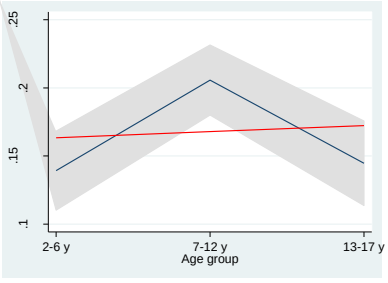
equal proportions?

```
. prop bullied, over(sex)
```


Ordered categories, trend

Does bullied vary with age?

```
twoway (fpfitci bullied agegr) ///
       (lfit    bullied agegr)
```



Ordered categories, trend

```
. tabulate bullied agegr,nofreq col chi2
```

Is bullied	Age group: 2-6 y	7-12 y	13-17 y	Total
no	86.08	79.43	85.53	83.23
yes	13.92	20.57	14.47	16.77
Total	100.00	100.00	100.00	100.00

Pearson chi2(2) = 14.1205 Pr = 0.001

equal proportions?

Trend?

```
. nptrend bullied, by(agegr)          /* Non-parametric
```

agegr	score	obs	sum of ranks
1	1	632	611892
2	2	807	834742.5
3	3	553	538393.5

z = 0.41
Prob > |z| = 0.683

Table of tests

	Numerical data		Proportions
	Normal	**Skewed**	
1 sample	One sample T-test	Kolmogorov-Smirnov	Binomial
2 independent samples	Independent sample T-test	Mann-Whitney U	Chi-square
K independent samples	ANOVA	Kruskal-Wallis	Chi-square
2 dependent samples	Paired sample T-test	Wilcoxon signed rank test	Mc-Nemar (2x2)

Categorical ordered:	use nonparametric tests