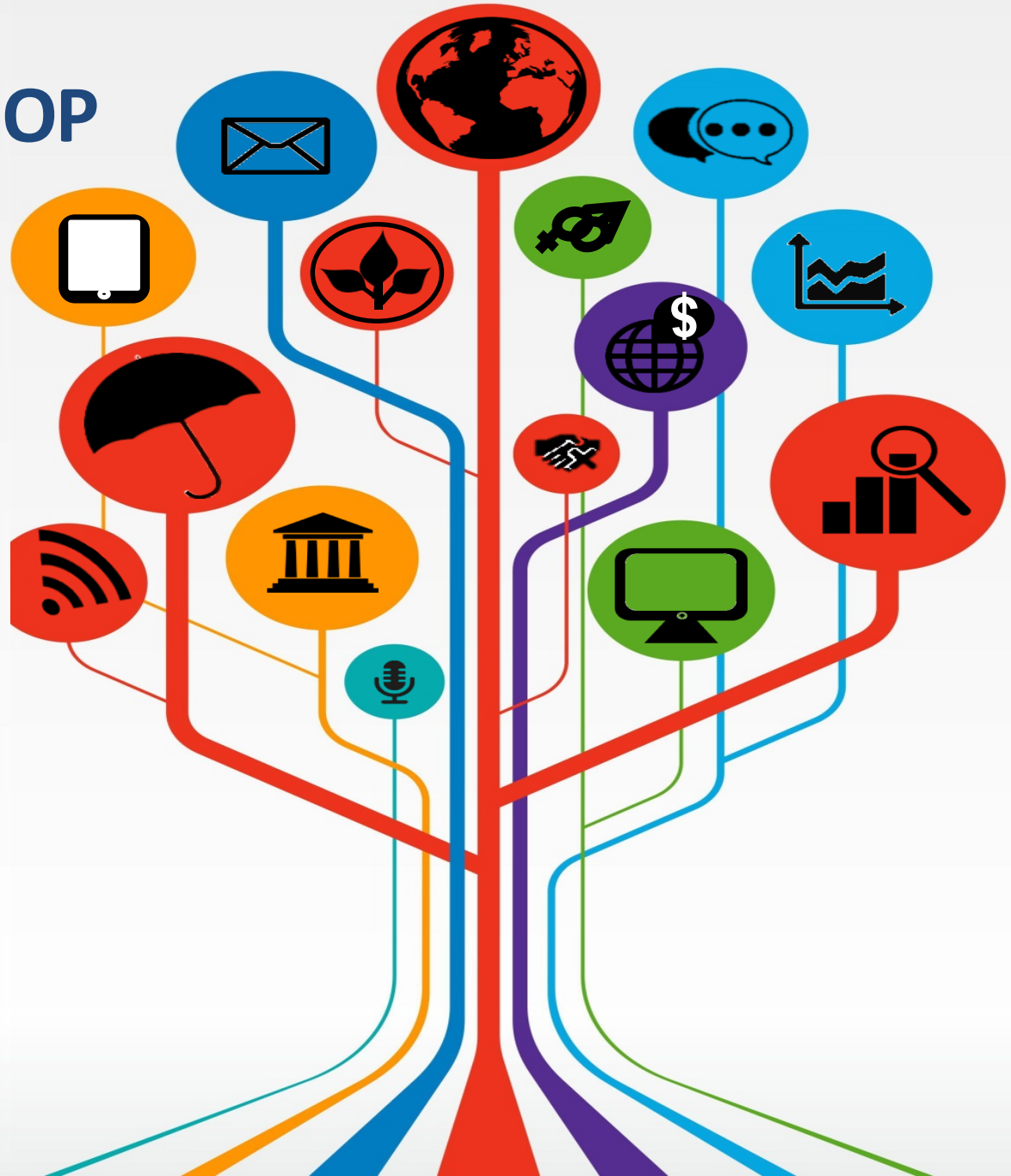
FIELD COORDINATOR WORKSHOP

Manage Successful Impact Evaluations

18 - 22 JUNE 2018
WASHINGTON, DC



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

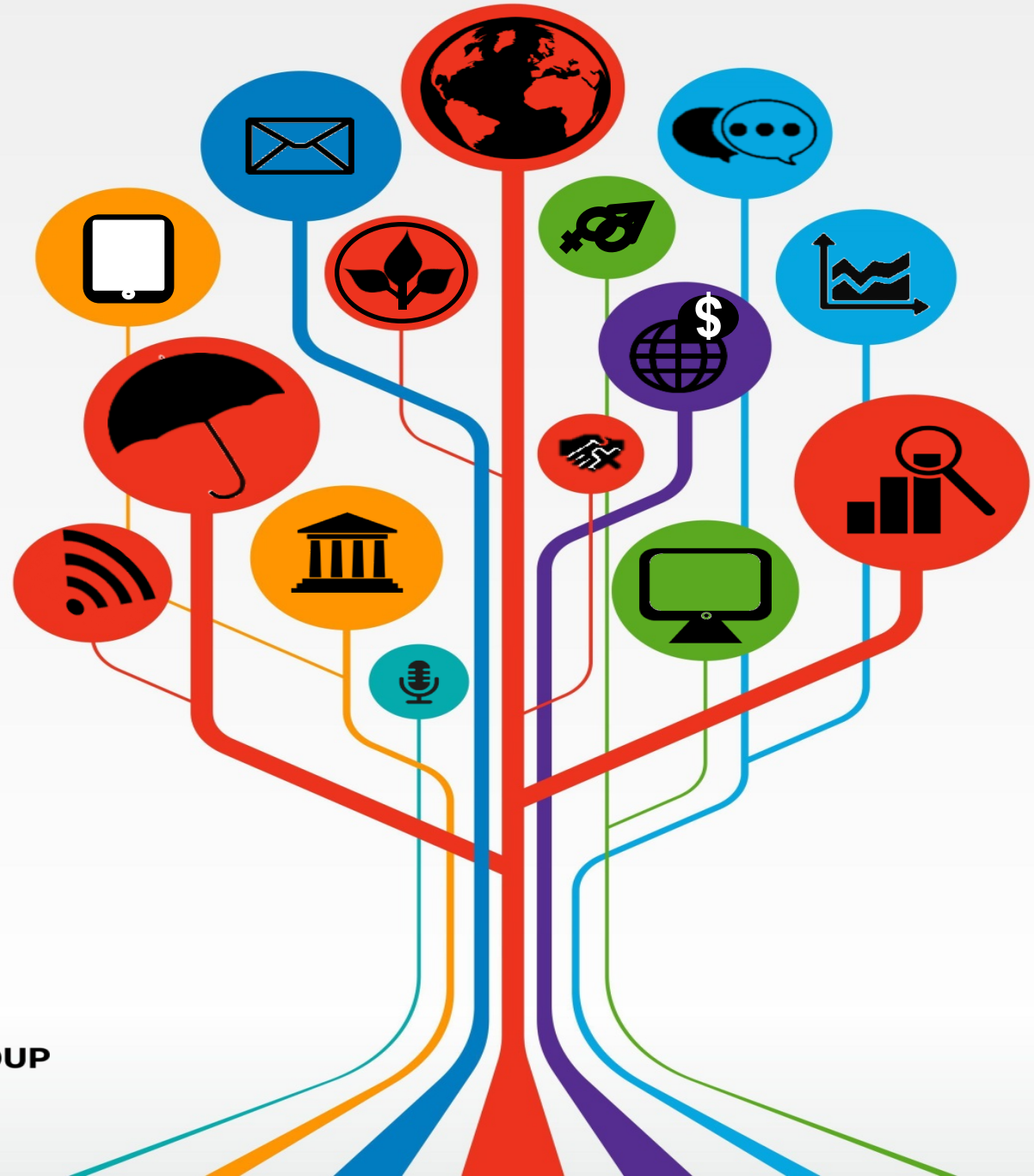
Descriptive Statistics: Creating Tables

Stata Track 2

Prepared by **DIME Analytics**
DIMEAnalytics_Internal_Use_Only@worldbank.org

Presented by **Benjamin Daniels and Sakina Shibuya**
bdaniels@worldbank.org | *sshibuya@worldbank.org*

http://www.worldbank.org/en/research/dime
June 21, 2018

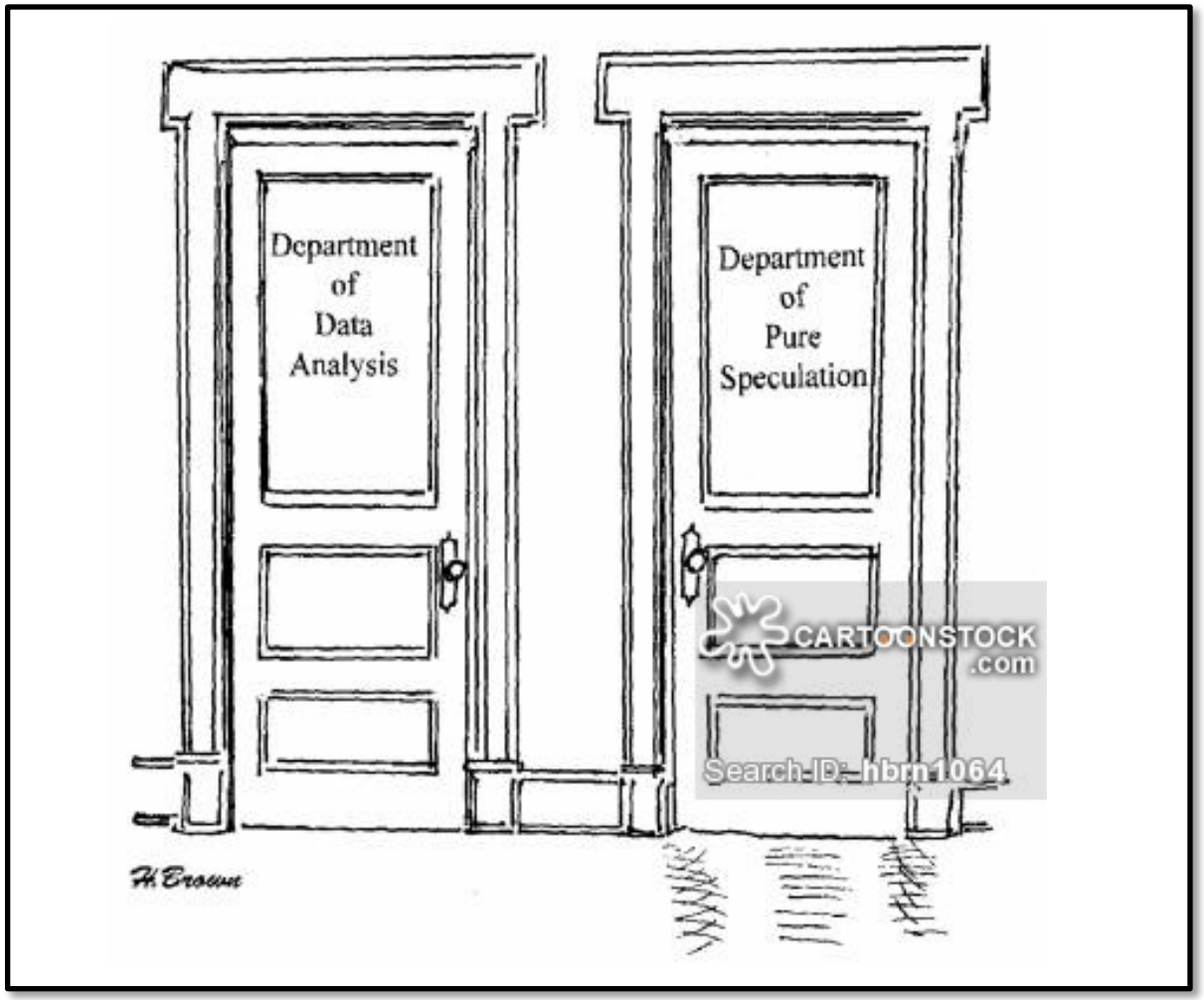


i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

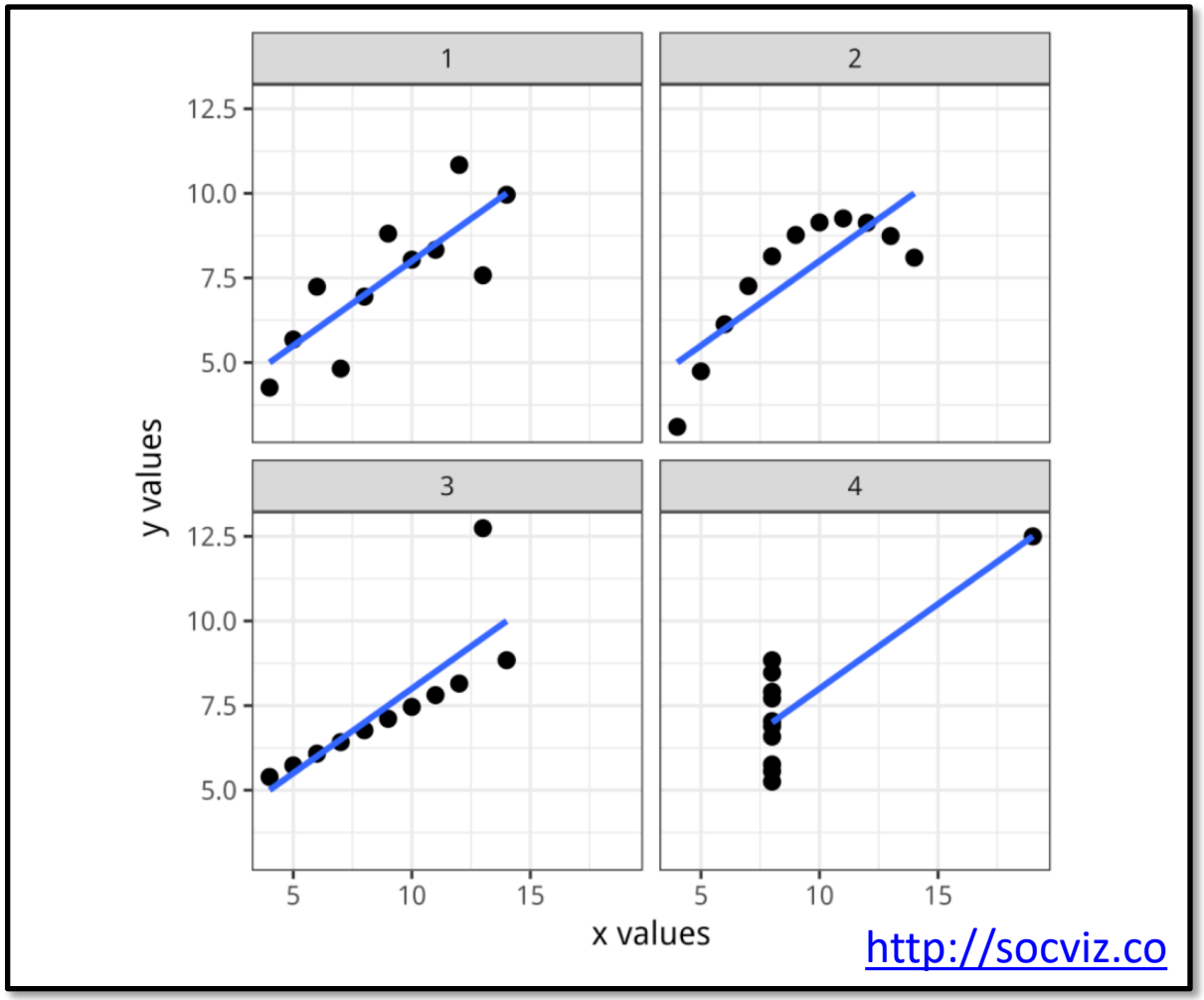
What are descriptive statistics?

- Numbers or figures that paint a picture of what a given dataset looks like
- They begin to help us understand the important features of our dataset, and can be useful in directing us towards areas of further analysis
- We will also show how to make basic regression outputs

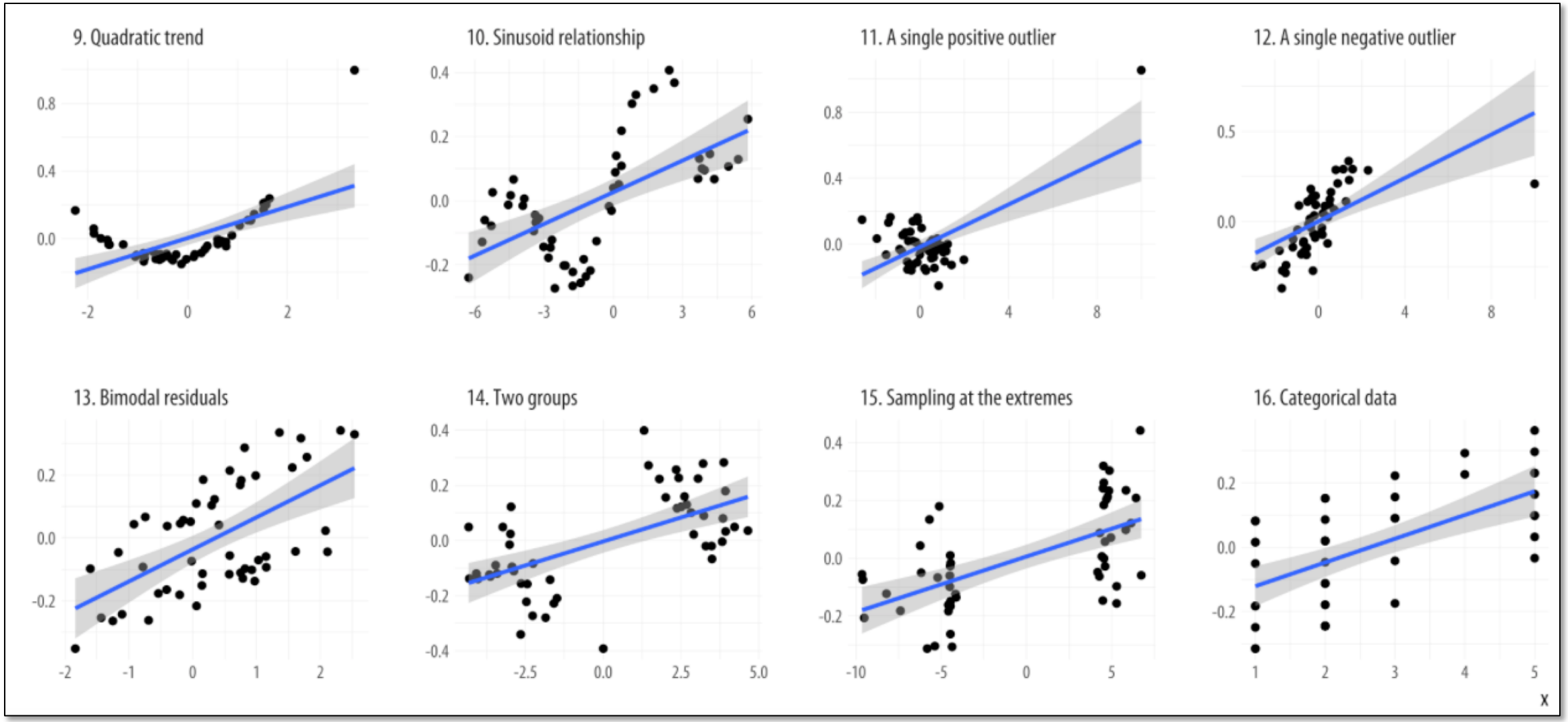


Descriptive statistics are NOT regressions

- This is “Anscombe’s Quartet”
- Every set here has:
  - The same means (x and y)
  - The same variances (x and y)
  - The same correlation coefficient
  - The same regression coefficient
  - The same regression R^2
- Regression analysis tells you *nothing* about the world if you don’t understand the shape of the world you are in!

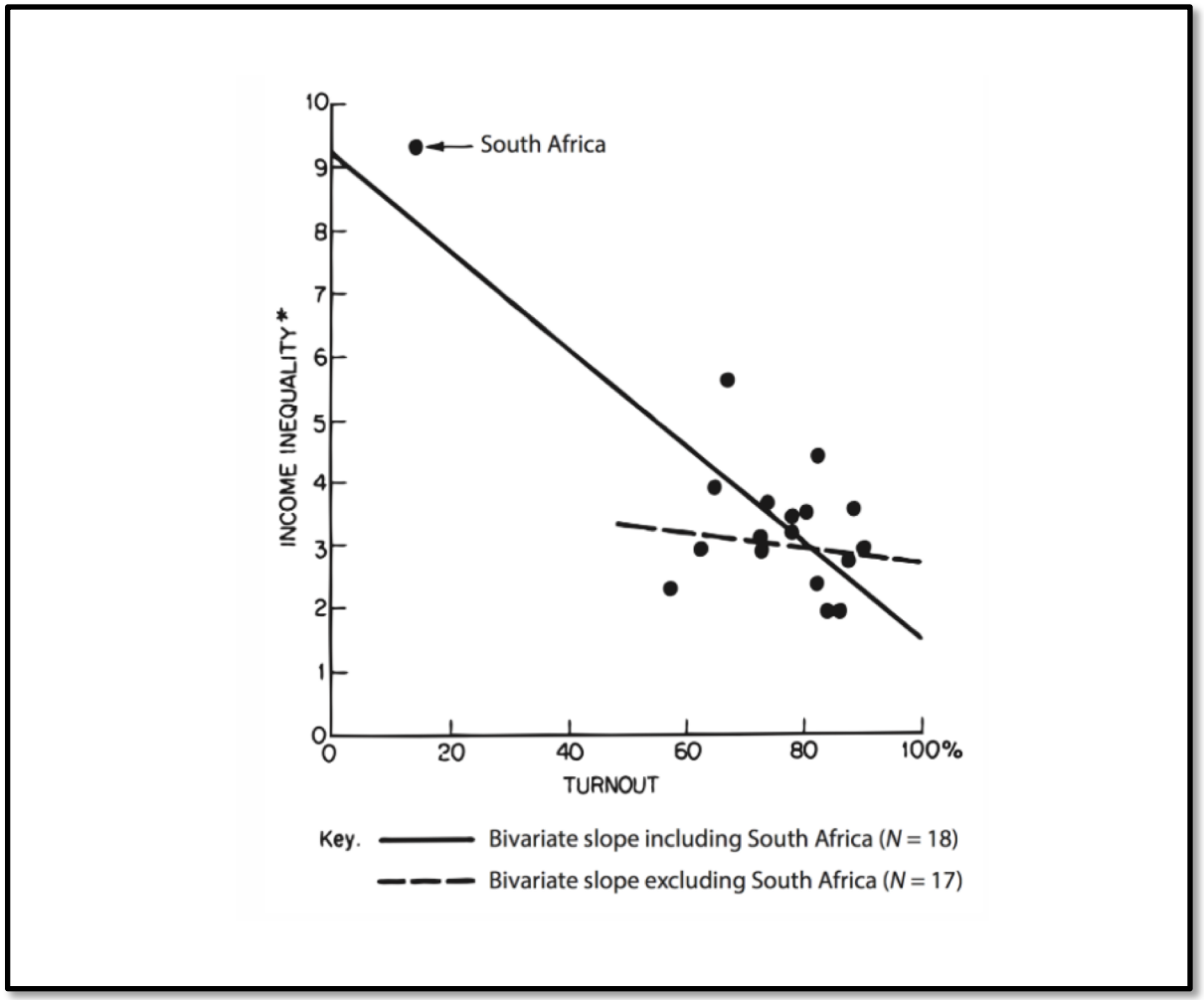


Data can take almost any shape



This really matters for impact analysis

- In this case, for example, simply running a regression on the data would give a very wrong impression about the strength of the relationship
- And real data has many more than two dimensions!



Tables present a lot of data in a small space

	All cities (Delhi, Mumbai, and Patna)		Patna and Mumbai only	
	Case 1	Case 2	Case 1	Case 2
Number of interactions	599	601	548	548
Referral	96, 0·16 (0·13–0·19)	401, 0·67 (0·63–0·70)	75, 0·14 (0·11–0·17)	362, 0·66 (0·62–0·70)
Ideal case management	80, 0·13 (0·11–0·16)	372, 0·62 (0·58–0·66)	64, 0·12 (0·09–0·14)	335, 0·61 (0·57–0·65)
Drugs				
Number of drugs	2·09 (1·99–2·20)	0·98 (0·88–1·09)	2·07 (1·97–2·18)	0·97 (0·86–1·08)
Antibiotic	221, 0·37 (0·33–0·41)	98, 0·16 (0·13–0·19)	200, 0·36 (0·32–0·41)	88, 0·16 (0·13–0·19)
Steroid	45, 0·08 (0·05–0·10)	16, 0·03 (0·01–0·04)	37, 0·07 (0·05–0·09)	13, 0·02 (0·01–0·04)
Antibiotic or steroid	230, 0·38 (0·34–0·42)	104, 0·17 (0·14–0·20)	208, 0·38 (0·34–0·42)	94, 0·17 (0·14–0·20)
Fluoroquinolone	61, 0·10 (0·08–0·13)	23, 0·04 (0·02–0·05)	61, 0·11 (0·08–0·14)	23, 0·04 (0·03–0·06)
Schedule H	401, 0·67 (0·63–0·71)	188, 0·31 (0·28–0·35)	367, 0·67 (0·63–0·71)	172, 0·31 (0·27–0·35)
Schedule H1	37, 0·06 (0·04–0·08)	19, 0·03 (0·02–0·05)	31, 0·06 (0·04–0·08)	16, 0·03 (0·02–0·04)
Schedule X	0	0	0	0
Anti-tuberculosis	0	0	0	0

Data are n, proportion (95% CI) or mean (95% CI).

Table 2: **Management of Case 1 and Case 2 for all cities and for Patna and Mumbai only**

Table 5. Trust in Outsiders After the Earthquake – Aid Robustness

	(1) Foreigners in General	(2) Westerners (Europeans & Americans)	(3) Own Village	(4) Extended Family	(5) Own Biradari/Qaum	(6) Own Region	(7) Ability to Work Together	(8) Difference: Westerners & Own Region
Number of observations	4,610	4,610	4,610	4,610	4,610	4,610	4,610	4,610
Mean	0.459	0.482	0.247	0.457	0.299	0.172	0.398	0.310
SD	0.498	0.500	0.431	0.498	0.458	0.377	0.490	0.567
PANEL A: Army Aid								
Distance to Fault (km)	-0.007***	-0.006***	0.001	0.002	0.001	-0.001	-0.003*	-0.005**
	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
Fraction Village Reporting Army Aid	-0.022	-0.052	0.062	0.069	0.014	-0.016	0.090	-0.036
	(0.087)	(0.088)	(0.067)	(0.094)	(0.079)	(0.053)	(0.078)	(0.081)
R^2	0.116	0.107	0.035	0.054	0.038	0.034	0.081	0.084
PANEL B: Western Aid								
Distance to Fault (km)	-0.003*	-0.003	0.000	0.002	0.001	-0.000	-0.001	-0.002
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Fraction Village Reporting Western Aid	0.375***	0.324***	0.028	0.048	0.007	0.085*	0.236***	0.238***
	(0.062)	(0.060)	(0.055)	(0.056)	(0.064)	(0.045)	(0.057)	(0.071)
R^2	0.141	0.125	0.034	0.054	0.038	0.036	0.090	0.092
PANEL C: Other Aid								
Distance to Fault (km)	-0.007***	-0.006***	-0.000	0.001	0.000	-0.001	-0.003**	-0.005**
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Fraction Village Reporting Other Aid	-0.066	-0.069	-0.058	-0.155**	-0.105	0.004	0.079	-0.073
	(0.083)	(0.091)	(0.057)	(0.075)	(0.075)	(0.052)	(0.087)	(0.101)
R^2	0.117	0.108	0.035	0.057	0.040	0.034	0.081	0.085
PANEL D: All Aid								
Distance to Fault (km)	-0.003*	-0.003	0.001	0.002	0.000	-0.000	0.000	-0.003
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Fraction Village Reporting Army Aid	-0.033	-0.066	0.053	0.038	-0.008	-0.015	0.113	-0.051
	(0.084)	(0.084)	(0.067)	(0.092)	(0.079)	(0.053)	(0.073)	(0.080)
Fraction Village Reporting Western Aid	0.380***	0.325***	0.018	0.020	-0.013	0.090*	0.263***	0.235***
	(0.062)	(0.060)	(0.057)	(0.059)	(0.067)	(0.050)	(0.061)	(0.073)
Fraction Village Reporting Other Aid	0.021	-0.002	-0.044	-0.143*	-0.109	0.024	0.165*	-0.025
	(0.076)	(0.086)	(0.062)	(0.077)	(0.081)	(0.057)	(0.086)	(0.101)
R^2	0.141	0.126	0.035	0.057	0.040	0.036	0.095	0.092

Notes: Estimated coefficients are reported from specifications that correlate trust with distance to the fault and aid controls. Panel A adds controls for the fraction of the village that reported aid from the Pakistan Army, Panel B for western organizations other than the U.N., Panel C for all other sources, and Panel D as a single combined regression. Each regression in columns 1-6 has as its dependent variable the binary-coded answer to the trust question. The dependent variable in column 7 is the working together question and in column 8 the difference between the answers to the trust questions from columns 2 and 6. Regressions also include controls for education, asset tercile, gender, distance to epicenter, and local slope as in Table 3 and 4. Standard errors in parentheses, clustered at the village level. (*** = p<0.01, ** = p<0.05, * = p<0.1)

43

Tables are important and also hard

- I will not focus on design here: putting tables together is hard enough!
- There are several options for both Excel and LaTeX

Creating print-ready tables in Stata

Michael Lokshin
The World Bank
Washington, DC
mlokshin@worldbank.org

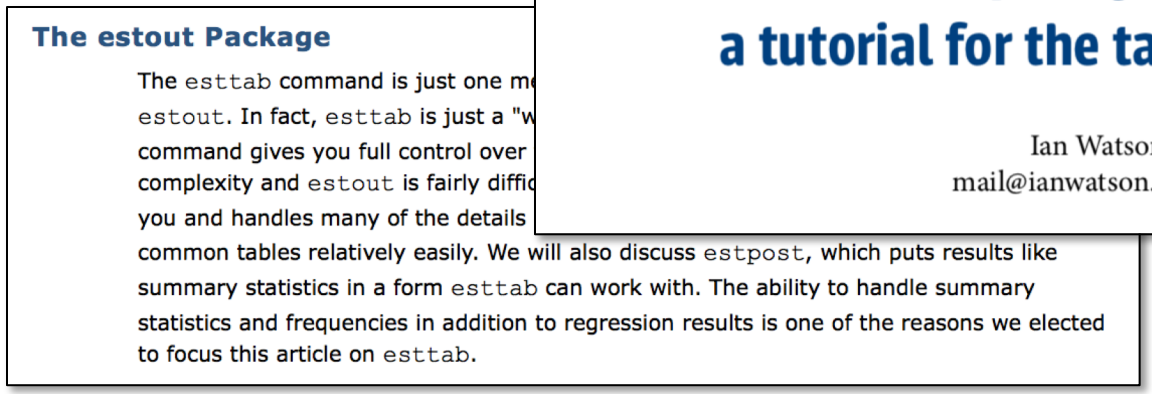
Zurab Sajaia
The World Bank
Washington, DC
zsajaia@worldbank.org

Abstract. This article describes the new Stata command `xml_tab`, which outputs the results of estimation commands and Stata matrices directly into tables in XML format. The XML files can be opened with Microsoft Excel or OpenOffice Calc, or they can be linked with Microsoft Word files. By using XML, `xml_tab` allows Stata users to apply a rich set of formatting options to the elements of output tables.

mates, regression, matrices, xml, Excel, Word

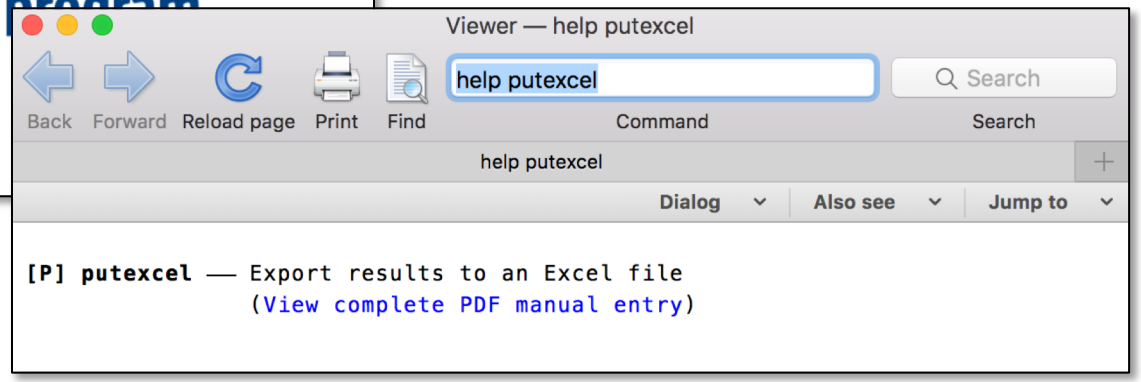
Publication quality tables in Stata:
a tutorial for the tabout program

Ian Watson
mail@ianwatson.com.au



The estout Package

The esttab command is just one m
estout. In fact, esttab is just a "w
command gives you full control over
complexity and estout is fairly diffic
you and handles many of the details
common tables relatively easily. We will also discuss estpost, which puts results like summary statistics in a form esttab can work with. The ability to handle summary statistics and frequencies in addition to regression results is one of the reasons we elected to focus this article on esttab.



Viewer — help putexcel

Back Forward Reload page Print Find help putexcel Command Search Search

help putexcel

Dialog Also see Jump to

[P] putexcel —— Export results to an Excel file
(View complete PDF manual entry)

Process reminder: Keep “raw” files separate

- Each table should be created into a *separate output file* that says exactly what it is, and have *its own section of code* to re-create
- It’s tempting to write impressive code to make all the tables at once in one file!
- But this makes your code *less* modular and readable

Three most common types of tables

- **Summary statistics**
  - Show an overview of variable distributions, possible for multiple groups
- **Balance tests**
  - Show a direct comparison of variable means across treatment arms
- **Regression outputs**
  - Estimate parameters of interest like treatment effects

[*sumStats*]

https://github.com/worldbank/stata/tree/master/src/sumStats

[*iebaltab*]

https://worldbank.github.io/ietoolkit

[*xml_tab*]

https://ideas.repec.org/c/boc/bocode/s456760.html

Summary statistics with [*sumStats*]

- [*sumStats*] is a command that will output anything you can get from [*tabstat*]

statname	Definition	*statname*	Definition
mean	mean	p1	1st percentile
count	count of nonmissing observations	p5	5th percentile
n	same as count	p10	10th percentile
sum	sum	p25	25th percentile
max	maximum	median	median (same as p50)
min	minimum	p50	50th percentile (same as median)
range	range = max − min	p75	75th percentile
sd	standard deviation	p90	90th percentile
variance	variance	p95	95th percentile
cv	coefficient of variation (sd/mean)	p99	99th percentile
semean	standard error of mean (sd/$\sqrt{n}$)	iqr	interquartile range = p75 − p25
skewness	skewness	q	equivalent to specifying p25 p50 p75
kurtosis	kurtosis		

- It also allows multiple [*if*]-restrictions with different variable lists

sumStats

sumStats will produce requested statistics for any number and combination of variables and sample restrictions.

	A	B	C	D	E	F
1		mean	sd	p5	p95	N
2	Price	6,072.423	3,097.104	3,667.000	13,594.000	52.000
3	Mileage (mpg)	19.827	4.743	14.000	29.000	52.000
4	Repair Record 1978	3.021	0.838	2.000	4.000	48.000
5	Headroom (in.)	3.154	0.916	1.500	4.500	52.000
6	Trunk space (cu. ft.)	14.750	4.306	7.000	21.000	52.000
7	Price	6,384.682	2,621.915	3,798.000	11,995.000	22.000
8	Mileage (mpg)	24.773	6.611	17.000	35.000	22.000
9	Repair Record 1978	4.286	0.717	3.000	5.000	21.000
10	Headroom (in.)	2.614	0.486	2.000	3.500	22.000
11	Trunk space (cu. ft.)	11.409	3.217	6.000	16.000	22.000

```
wb_git_install sumStats
sysuse auto , clear
sumStats ///
  (price mpg rep78 headroom trunk if foreign == 0) ///
  (price mpg rep78 headroom trunk if foreign == 1) ///
  using "table_1.xls" ///
  , replace stats(mean sd p5 p95 N)
```

Multiple levels or groups are easy

- Village statistics can be called with [*if tag_village == 1*]
- Treatment group can be called with [*if treatment == 1*]
- And so on, with only one line of code in Stata

One line of well-formatted code

Table 1. Descriptive Statistics

	mean	sd	p25	p50	p75	N
Distance to Fault (km)	17.477	14.149	5.555	13.564	24.311	28,297.000
Distance to Epicenter (km)	36.373	17.490	25.115	35.161	48.013	28,297.000
Closest Faultline (km)	2.799	2.486	0.773	1.984	4.143	28,297.000
Death in HH During Quake	0.061	0.240	0.000	0.000	0.000	28,297.000
Home Destroyed	0.572	0.495	0.000	1.000	1.000	8,351.000
Home Damaged or Destroyed	0.911	0.285	1.000	1.000	1.000	8,350.000
Household Size	5.477	2.332	4.000	5.000	7.000	28,297.000
Total Annual Food Expenditu	83,207.844	88,160.971	37,500.000	62,280.000	98,805.000	2,456.000
Total Annual Nonfood Expen	84,207.025	109,511.091	26,786.500	46,182.500	93,035.000	2,456.000
Abbotabad	0.206	0.405	0.000	0.000	0.000	2,456.000
Bagh	0.175	0.380	0.000	0.000	0.000	2,456.000
Mansehra	0.276	0.447	0.000	0.000	1.000	2,456.000
Muzaffarabad	0.342	0.475	0.000	0.000	1.000	2,456.000
Family Size	6.123	2.689	4.000	6.000	8.000	2,455.000
Asset Index (PCA) (Pre-Quak	0.002	0.999	-0.551	-0.093	0.568	2,456.000
House Destroyed in Quake?	0.599	0.490	0.000	1.000	1.000	2,455.000
Eligible for death compensatic	0.149	0.433	0.000	0.000	0.000	2,455.000
Eligible for housing compensa	0.925	0.320	1.000	1.000	1.000	2,455.000
Eligible for injury compensati	0.156	0.438	0.000	0.000	0.000	2,455.000
Eligible for lcgs compensation	0.886	0.710	0.000	1.000	1.000	2,455.000
Private School Binary (Post-Q	0.217	0.412	0.000	0.000	0.000	3,089.000

```
sumstats ///
  (hh_faultdist hh_epidist hh_fault_minimum c_death_quake c_home_des c_home_dam hh_members if tag_hh == 1) /// Census hh data
  (hh_consumption_food hh_consumption_nonfood ///
    hh_district_1 hh_district_2 hh_district_3 hh_district_4 ///
    hh_familysize hh_assets_pca_pre hh_stats_destroyed hh_*_eligible hh_n_children_u6 hh_head_female ///
    hh_aid_any hh_aid_any_1 check if tag_hh == 1 & touse_shock == 1) /// Household Survey Data
  (hh_slope if tag_uc == 1) /// UC Data
  (indiv_male indiv_age u15 if indiv_dead == 0 ) ///
  (indiv_agecat_2 indiv_agecat_3 indiv_agecat_1 ///
    indiv_father_edu m_indiv_edu_binary m_indiv_age m_indiv_health_height ///
    indiv_health_height indiv_health_weight indiv_school_enrolled_post indiv_school_pri_bi_post ///
    if indiv_dead == 0 & touse_shock == 1 & indiv_age > 2 & indiv_age < 16) /// Children survey data
  ///
  using "$directory/Outputs/Shock/1_descriptives.xls" ///
,   replace title("Table 1. Descriptive Statistics") sheet("Table 1") ///
  stats(mean sd p25 p50 p75 N) lines(COL_NAMES 3 LAST_ROW 3)
```

Balance tables with [*iebaltab*]

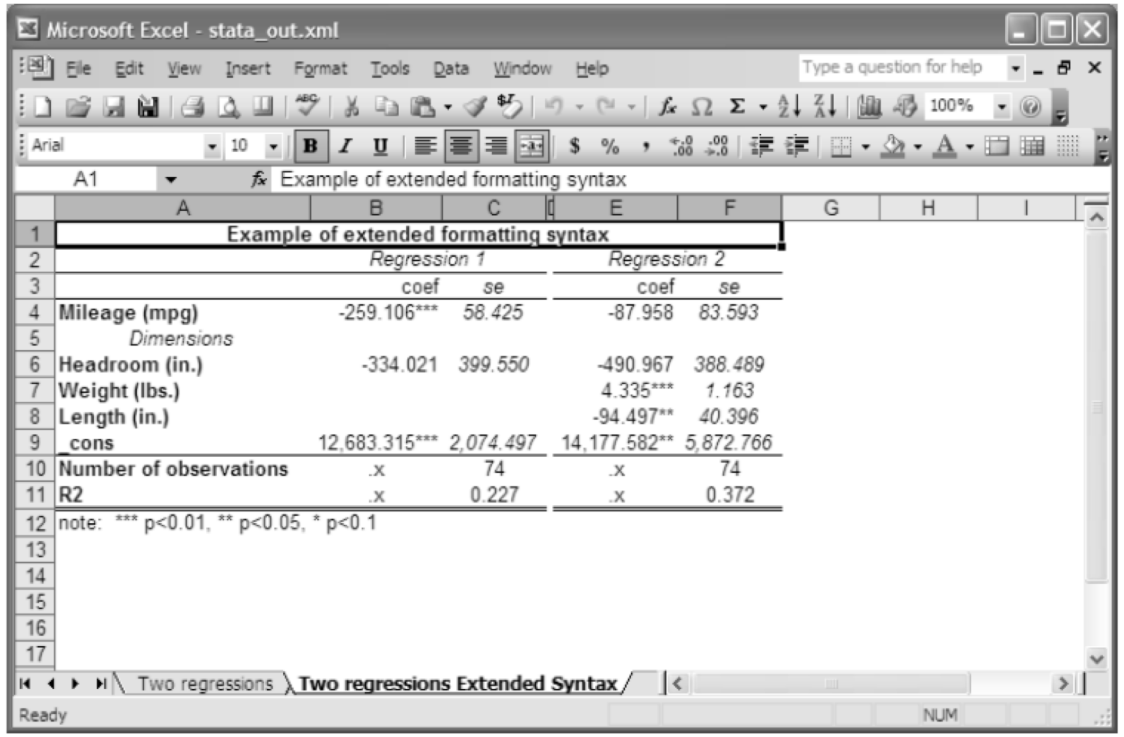
- Balance tables feature in almost every impact evaluation
- We use balance tables to show that there was no difference between our control and treatment group in the baseline before the intervention
- To us [*iebaltab*], list all the variables you want to test balance in, and use the option [*grpvar*()] to indicate which group each observation is in.

Variable	(1) Control Mean/SE	(2) Treatment Mean/SE	T-test Difference (1)-(2)
Age in years	42.880	42.126	0.754
	(1.746)	(0.535)	
Respondent is male	0.538	0.479	0.059
	(0.050)	(0.008)	
Years of schooling	10.930	10.838	0.092
	(0.171)	(0.183)	
Respondent is employed	0.835	0.892	-0.057
	(0.060)	(0.041)	
Monthly earnings (number of minimum wages)	1.582	1.491	0.091
	(0.094)	(0.067)	
Average commuting distance	18.241	11.737	6.504***
	(1.078)	(0.233)	
N	158	167	
Clusters	6	6	
F-test of joint significance (F-stat)			9.892***

```
iebaltab    age d_male educ d_employed earnings distance, ///
            covariates(stratum) ///
            grpvar(tmt_status) ///
            vce(cluster neighborhood) ///
            savetex("$outputs/balance_table") ///
            replace onenrow ftest rowvarlabel
```

Outputting regressions with [*xml_tab*]

- **Powerful regression engine:** automatically aligns regression coefficients across models
- **Accepts arbitrary matrices:** Very easy to hack [*xml_tab*] into doing whatever you want
- **Flexible formatting syntax:** Allows reasonable customization of decimal places, stars, etc. via Excel formatting styles



		Regression 1		Regression 2	
		coef	se	coef	se
Mileage (mpg)		-259.106***	58.425	-87.958	83.593
	Dimensions				
Headroom (in.)		-334.021	399.550	-490.967	388.489
Weight (lbs.)				4.335***	1.163
Length (in.)				-94.497**	40.396
_cons		12,683.315***	2,074.497	14,177.582**	5,872.766
Number of observations		.x	74	.x	74
R2		.x	0.227	.x	0.372

note: *** p<0.01, ** p<0.05, * p<0.1

Figure 5. Output using extended formatting options

Regressions are automatically stacked up

```
qui {

  reg m_indiv_edu_binary m_eligible_2 `mother_controls' hh_faultdist , cl(village_code)
    estimates store fal1
    test m_eligible_2
    local f = round(r(F),.01)
    estadd scalar f = `f'

  xi: reg m_indiv_edu_binary m_eligible_2 `mother_controls' hh_faultdist if m_indiv_momedu_false8≠0 , cl(village_code)
    estimates store fal2
    test m_eligible_2
    local f = round(r(F),.01)
    estadd scalar f = `f'

  reg m_indiv_edu_binary m_eligible_2 `fault_controls' `mother_controls' hh_faultdist , cl(village_code)
    estimates store fal3
    test m_eligible_2
    local f = round(r(F),.01)
    estadd scalar f = `f'

  reg m_indiv_edu_binary m_indiv_sb8 `fault_controls' `mother_controls' hh_faultdist hh_logcons , cl(village_code)
    estimates store fal4
    test m_indiv_sb8
    local f = round(r(F),.01)
    estadd scalar f = `f'

  reg m_indiv_edu_binary m_eligible_2 m_eligible_3 m_eligible_4 `fault_controls' `mother_controls' hh_faultdist hh_logcons if instrument≠., cl(village_code)
    estimates store fal5
    test m_eligible_2
    local f = round(r(F),.01)
    estadd scalar f = `f'

  }

  gen f = 0
    label var f "F-statistic for Age 9 School Availability" // labelling for output

    xml_tab fal1 fal2 fal3 fal4 fal5 ///
    using "$directory/Outputs/Shock/4_instrument.xls" ///
    , title("Table 4. Instrument Falsification Tests and First Stage F-tests (Dependent variable: Probability of completing primary school)") ///
      replace below cnames("Instrument" "Recieved School Sometime" "Geographical Controls" "Boys' School" "Girls' School (Other Ages)") c("Constant") ///
      showeq ceq($numbering) stats(N f) format((S2110) (SCCB0 N2303)) lines(COL_NAMES 3 LAST_ROW 3) ///
      keep( hh_faultdist m_eligible_2 m_indiv_sb8 m_eligible_3 m_eligible_4  _cons) drop(o.*) ///
      note("Controlled for individual and geographical characterics.", "Standard errors clustered by village.")
```

Table 4. Instrument Falsification Tests and First Stage F-tests (Dependent variable: Probability of completing primary school)

	fal1	fal2	fal3	fal4	fal5
	(1)	(2)	(3)	(4)	(5)
	Instrument	Recieved School Sometime	Geographical Controls	Boys' School	Girls' School (Other Ages)
Distance from Faultline (km)	0.001	0.001	-0.000	-0.000	-0.001
	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)
Girls' school present by age 9	0.143***	0.141***	0.131***		0.151***
	(0.031)	(0.033)	(0.030)		(0.043)
Boys' school present by age 8				-0.010	
				(0.034)	
Girls' school present at age 10-14					0.035
					(0.055)
Girls' school present after age 14					0.022
					(0.035)
Constant	0.812***	-0.635***	0.333	-0.772	-0.590
	(0.104)	(0.180)	(0.254)	(0.486)	(0.474)
Number of observations	947	808	947	947	947
F-statistic for Age 9 School Availability	21.860	18.420	18.770	0.090	12.590
note: *** p<0.01, ** p<0.05, * p<0.1					
Controlled for individual and geographical characterics.					
Standard errors clustered by village.					

Helpful checklist before sending tables to PI

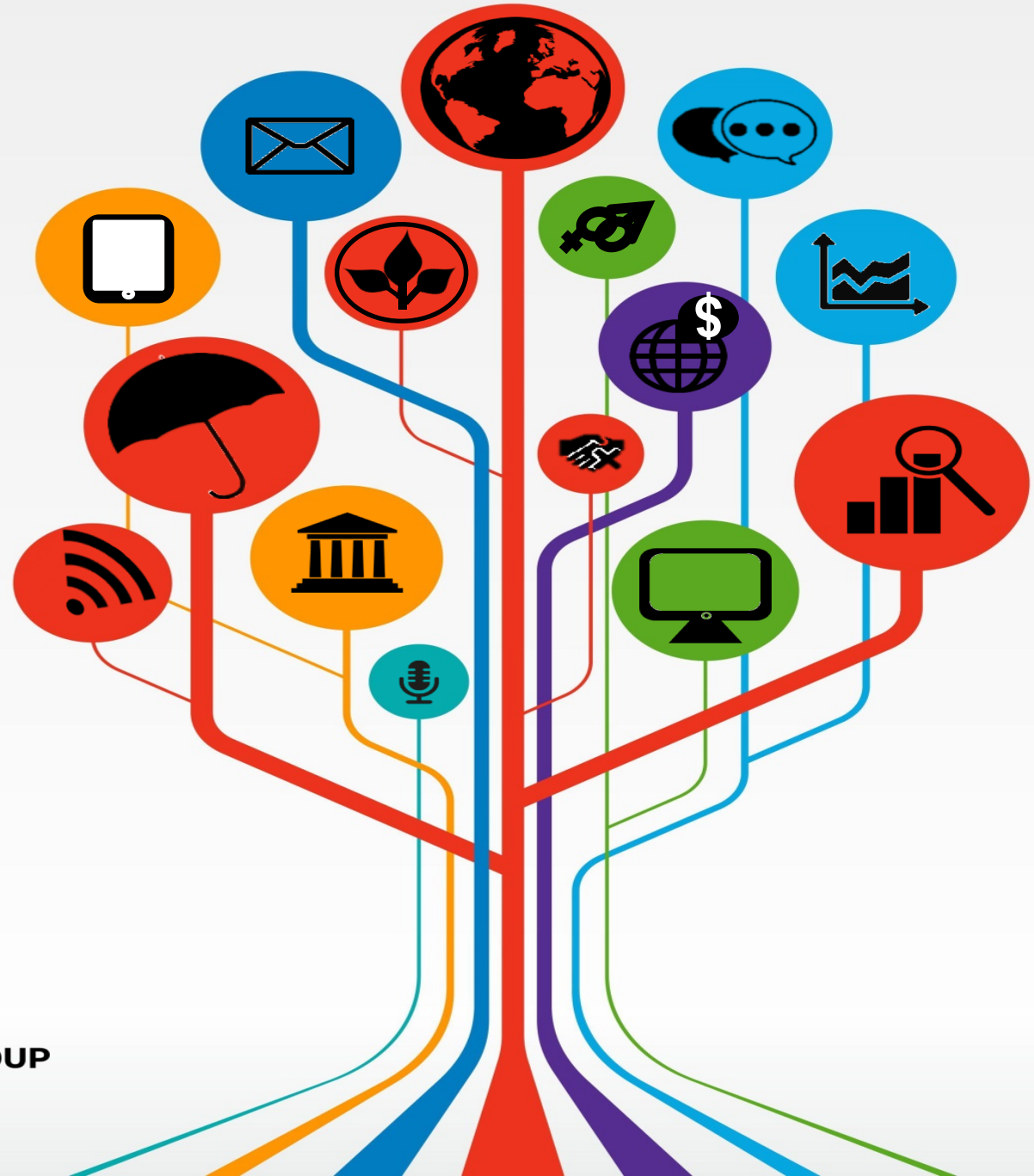
- Does the number of observations for each regression or summary statistic make sense?
- Do the magnitude and sign of each coefficient/summary statistic seem reasonable?
- Did you delete the constant term and add the control mean in the regression table?
- Did you check for joint significance of your covariates?
- Did you label the dependent variables/columns?
- Did you label the covariates/rows?

https://blogs.worldbank.org/impactevaluations/generating-regression-and-summary-statistics-tables-stata-checklist-and-code

- Did you add a title?
- Is it clear what the estimation procedure is (e.g. regression vs. probit)?
- Are the column widths the right size so as not to cut off text?
- Is the bordering consistent with your other tables?
- Are the numbers rounded to an appropriate level, so you don't display too many decimal places?
- Do the notes to the table clearly indicate how standard errors have been estimated, and what control variables if any have been included but not shown?

https://dimewiki.worldbank.org/wiki/Checklist:_Submit_Table_Checklist

Thank you!



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP