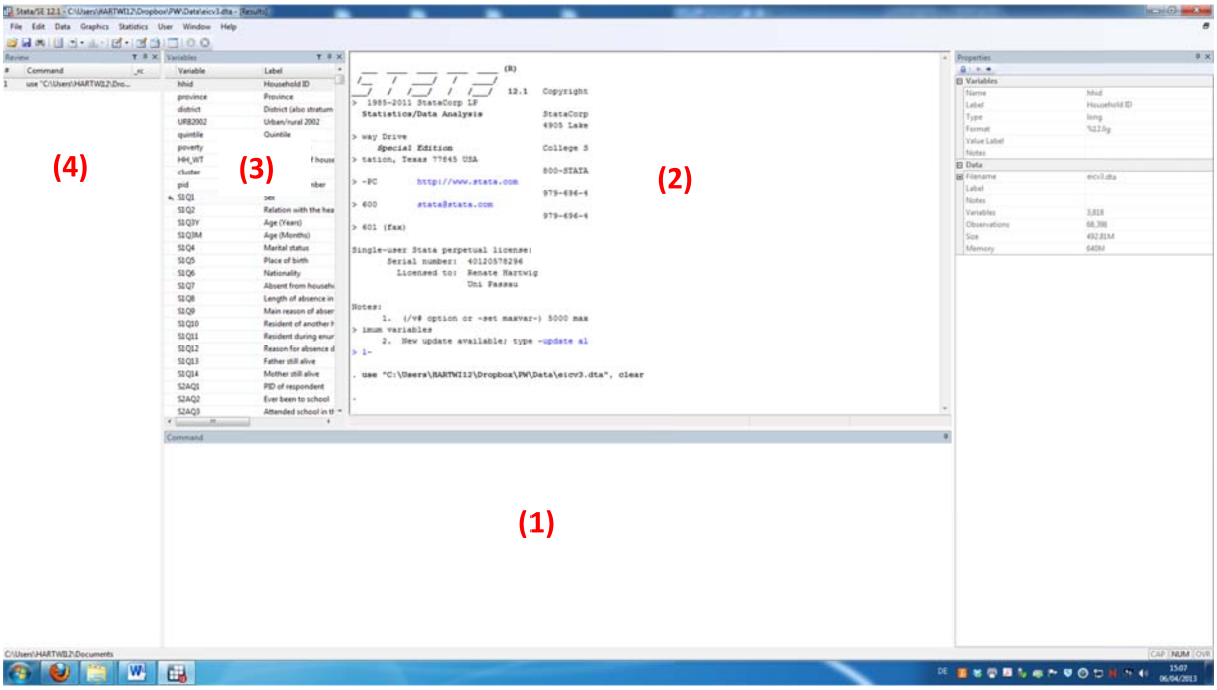
Introduction to Stata

In this course we will be working with Stata12.

1. The Stata Environment


Stata/SE 12.1 - C:\Users\HARTWI12\Dropbox\PW\Data\eicv3.dta - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Variables
Command
Properties
(4)
(3)
(2)
(1)
C:\Users\HARTWI12\Documents
CAP NUM OVR
15:07
06/04/2013

#	Command	_rc
1	use "C:\Users\HARTWI12\Dro...	

Variable	Label
hhid	Household ID
province	Province
district	District (also stratum
URB2002	Urban/rural 2002
quintile	Quintile
poverty	
HH_WT	f house
cluster	
pid	nber
S1Q1	sex
S1Q2	Relation with the hea
S1Q3Y	Age (Years)
S1Q3M	Age (Months)
S1Q4	Marital status
S1Q5	Place of birth
S1Q6	Nationality
S1Q7	Absent from househ
S1Q8	Length of absence in
S1Q9	Main reason of abser
S1Q10	Resident of another h
S1Q11	Resident during enur
S1Q12	Reason for absence d
S1Q13	Father still alive
S1Q14	Mother still alive
S2AQ1	PID of respondent
S2AQ2	Ever been to school
S2AQ3	Attended school in th

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   12.1   Copyright
>  1985-2011 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lake
> way Drive
     Special Edition                  College S
> tation, Texas 77845 USA
                                      800-STATA
> -PC        http://www.stata.com
                                      979-696-4
> 600        stata@stata.com
                                      979-696-4
> 601 (fax)

Single-user Stata perpetual license:
       Serial number:  40120578296
         Licensed to:  Renate Hartwig
                       Uni Passau

Notes:
      1.  (/v# option or -set maxvar-) 5000 max
> imum variables
      2.  New update available; type -update al
> l-

. use "C:\Users\HARTWI12\Dropbox\PW\Data\eicv3.dta", clear

.
```

Variables	
Name	hhid
Label	Household ID
Type	long
Format	%12.0g
Value Label	
Notes	
Data	
Filename	eicv3.dta
Label	
Notes	
Variables	3,818
Observations	68,398
Size	492.81M
Memory	640M

The 4 main windows are:

(1) The Command Window: where you enter/type commands
(2) The Results Window: where the results of a command are displayed
(3) The Variables Window: where the list of variables in your dataset are displayed
(4) The Review Window: which shows the trail of commands recently executed

2. Data Entry

There are several ways to read data into Stata's data editor. The data editor is a spreadsheet that looks very much like Excel.

For small datasets in Excel format the easiest way to load your data into Stata is to copy and paste it into the data editor:

- Arrange your data in Excel (variables across and observations down)
- Copy your data (incl. variable names) and go back to Stata
- Click on the data editor icon (8th Icon from the left) and paste (Ctrl+V)
- Save your file in Stata format (File-> Save as…)

If your data is already in Stata format (i.e. with the extension **.dta**), you have two options:

- Use the command window and type **use file location** (e.g. c:\India.dta)
- Click File->Open->(choose directory)

Data saved in other statistical packages like SPSS, SAS etc. can be converted into .dta-files using, for example, the software 'StatTransfer'.

3. Data Management

Once you load and save your data in Stata, the next step is to explore the nature of your data and start the analysis. If the data has been collected through a survey, reading the associated questionnaire in parallel obviously helps a lot to understand the content. Beyond that, however, there are still a few preparations to be made.

3.1 Log-File

Stata allows you to keep a running record of your statistical results – i.e. a separate record of the results of your commands. This is done in a log file. You can open a log file by:

- Clicking on the log-file icon (4th icon on the left)
- Clicking File->Log->Begin

A dialog box appears inviting you to type a file name and select a save-as type. Save your log-files in **.log** (not .smcl) as this allows you to edit your results later on.

If you want to temporarily stop recording your results (maybe because you are not sure you need to keep them), just type **log off** in the command window. Whenever you want to resume recording type **log on** in the command window. To close a log file type **log close**.

3.2 Do-File

Instead of typing each command individually in the command window, Stata allows you to execute a number of commands at once from a do file. A do-file works like a programme file. Do-files can be opened by:

- Clicking on the do-file icon (7th icon on the left)

An advantage of a do file is that you can repeat your sequence of commands as often as you want (or modified versions of it). It also allows you to share an analysis with others.

4. Basic Commands

In the following we use data from a labour force survey covering 261 workers in an industrial town in southern India in 1990.

4.1 Data Description

To see the content of your data click Data->Describe data->Describe data in memory

Alternatively you can also just type **describe** in the command window. This command displays the number of variables and observations in the data as well as the list of all variable names and labels.

In the following we will work with the commands directly instead of clicking through the menu (you can explore this option in your own time).

4.2 Summary Statistics

By summary statistics we refer to the mean, median, standard deviation, percentiles etc. It is important to differentiate between two groups of summary statistics: *mean based* or *order based* summary statistics.

Mean based summary statistics include the mean, variance, standard deviation, skewness, and kurtosis. The order based summary statistics include the median, the first and third quartiles, the inter quartile range (IQR). These statistics provide us with measures of central tendency (a value to which all observations converge), measures of spread (how observations are scattered around the central value) and also the shape of the distribution.

```
summarize age
sum age wage
sum wage, detail
```

Another useful command to get summary statistics is **tabstat**.

```
tabstat age
tabstat age, stats (n mean var sd cv p25 p50 p75 IQR)
tabstat age wage, stats (n mean var sd cv p25 p50 p75 IQR)
```

The preceding summary statistics are useful when we have quantitative variables. There is however a class of variables whose numeric values do not carry the usual meaning and property. We call such variables categorical variables. Categorical variables do not measure a concept as quantitative variables; they are rather count variables in the sense that we use them to count individual members with certain characteristics. In our data set *gender* and education are examples of categorical variables.

For categorical variables we do not use the usual measures of central tendency and spread, like the mean and standard deviation, simply because the number in categorical variables represent class/quality. Instead we aim to count the number of observations (frequency or relative frequency) in each category of a categorical variable.

```
tab education
tab education gender
tab education gender, row
tab education gender, column
tab education gender, sum(wage)
```

Stata commands can be qualified so as to operate over a specific range/portion of the data. You have seen that **tab education** produces the frequency table of *education* for all individuals in your sample. Suppose you want to see the frequency table of *education* only for women then you can use **if** as a qualifier:

```
tab education if gender==1
```

Similarly, you may be interested to see the age, gender and education levels of the top 10 wage earners in your sample:

```
sort wage
list wage in -10/l

list wage in 1/10
sum wage if age<30
list wage if gender==1 & wage<30
```

4.3 Creating New Variables

To create new variables we use the **generate** (short **gen**) command.

```
gen doublewage=wage*2
gen wage2=wage*wage
gen wage_2=wage^2
gen lnwage=ln(wage)

rename doublewage dblwage

replace wage2=wage/1000
```

In your data you have information on the age of the individual. Suppose you want to create a categorical variable with age groups.

```
gen agecat=1 if age<20
replace agecat=2 if age>=20 & age <40
replace agecat=3 if age>=40

tab agecat

label variable agecat “age groups 20 20-40 40+”
```

To delete variables from your data:

```
drop wage dblwage

keep wage dblwage
```

Note the difference!

5. Help and Find-it Function

5.1 Help Function

Stata has an online-based help-function. For example, if you are looking for more details on the commands and how they are specified type **help** followed by the command you are looking for into the command window.

e.g. **help sum**

5.2 Find-it Function

The find-it Function works similar to the help function and is particularly useful to look for programme packages that are not included in the standard Stata version.

e.g. **findit psmatch2**

References:
Kohler, U., and F. Kreuter (2005). *Data Analysis using Stata*. Texas: Stata Press.

Exercise:

- What is the average wage for males only? What is average female wage?
- How much of the wage can be explained by gender and education?
- Show the distribution of the education variable graphically.

Also, for further exercise and practise see Khandker *et al.* Chapter 11 (Introduction to Stata).

Practise with Stata

The file DHS2000.dta contains data on 11,926 children from the Demographic and Health Survey conducted in Malawi in 2000. The file DHS2005.dta contains information of 10,914 children from the Demographic and Health Survey in Malawi conducted in 2005.

Your tasks:

1. Load the data into Stata

2. Create a log file

3. Create a do file

4. Explore your data

    a. Of what age are the children in your data?

5. Combine both data sets into a single file named Malawi.dta

6. Create the following variables for analysis

    a. d_rural: a dummy indicating whether the child lives in a rural or urban environment.
    b. d_twin: a dummy indicating whether the child is a twin or not.
    c. d_male: a dummy indicating whether the child is male or not.
    d. d_dead: a dummy indicating whether the child has died or not.
    e. d_femalehead: a dummy indicating whether the head of the household is female or not.
    f. d_married: a dummy indicating whether the mother is married.
    g. mom_age_birth: a variable indicating the mother's age at the birth of the child.
    h. childage: a variable with the current age of the child in years.
    i. d_2005 a dummy indicating whether the child has been surveyed in 2004/5 or not.

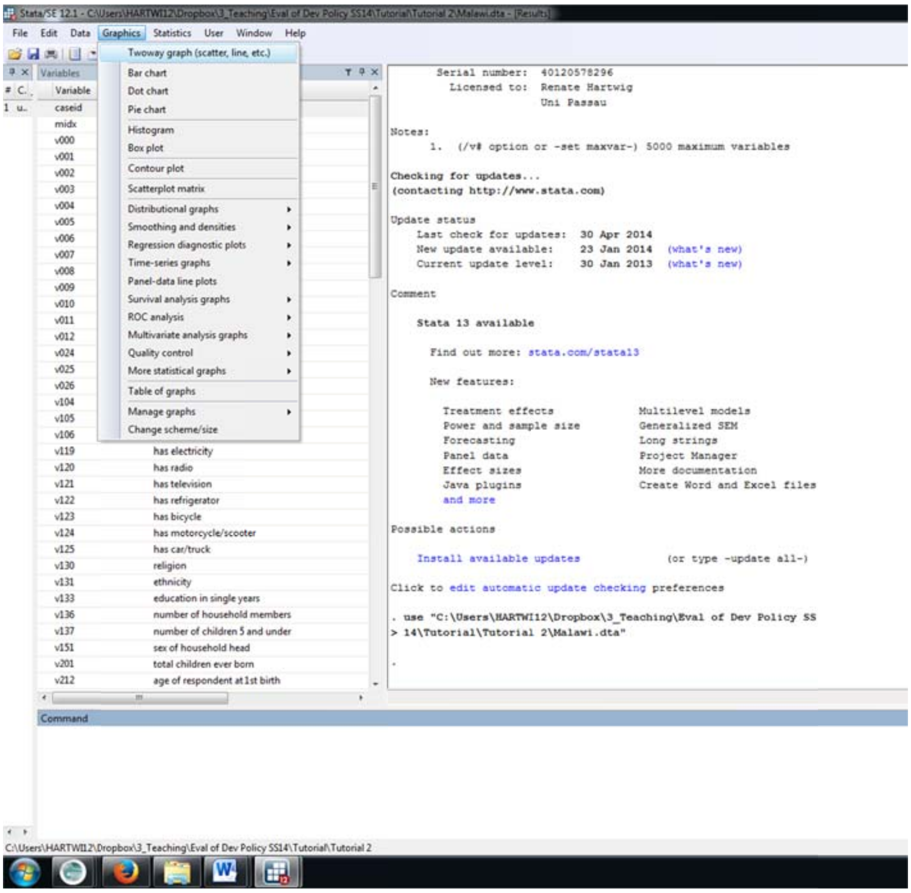
7. Compare the characteristics of the children in 2000 to those in 2005. Are there significant differences between them?

8. Show your results in a table.

Graphics with Stata

Load the Malawi dataset (Malawi.dta) that you have been using previously.

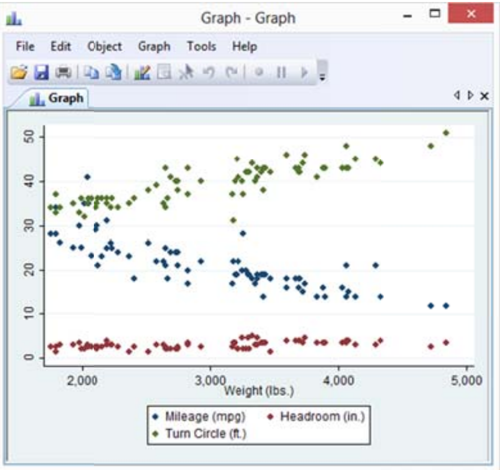
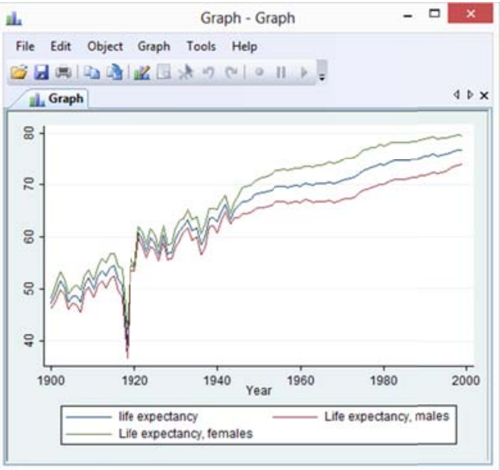
As a starting point, all basic graphs can be produced via the Menu ‘Graphics’



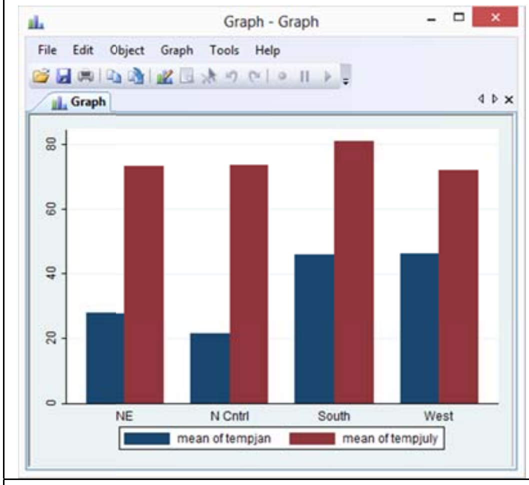
Variable	
caseid	
midx	
v000	
v001	
v002	
v003	
v004	
v005	
v006	
v007	
v008	
v009	
v010	
v011	
v012	
v024	
v025	
v026	
v104	
v105	
v106	
v119	has electricity
v120	has radio
v121	has television
v122	has refrigerator
v123	has bicycle
v124	has motorcycle/scooter
v125	has car/truck
v130	religion
v131	ethnicity
v133	education in single years
v136	number of household members
v137	number of children 5 and under
v151	sex of household head
v201	total children ever born
v212	age of respondent at 1st birth

```
        Serial number:  40120578296
          Licensed to:  Renate Hartwig
                        Uni Passau

Notes:
      1.  (/v# option or -set maxvar-) 5000 maximum variables

Checking for updates...
(contacting http://www.stata.com)

Update status
    Last check for updates:  30 Apr 2014
    New update available:    23 Jan 2014  (what's new)
    Current update level:    30 Jan 2013  (what's new)

Comment

    Stata 13 available

      Find out more: stata.com/stata13

      New features:

        Treatment effects             Multilevel models
        Power and sample size         Generalized SEM
        Forecasting                   Long strings
        Panel data                    Project Manager
        Effect sizes                  More documentation
        Java plugins                  Create Word and Excel files
        and more

Possible actions

    Install available updates             (or type -update all-)

Click to edit automatic update checking preferences

. use "C:\Users\HARTWI12\Dropbox\3_Teaching\Eval of Dev Policy SS
> 14\Tutorial\Tutorial 2\Malawi.dta"

.
```

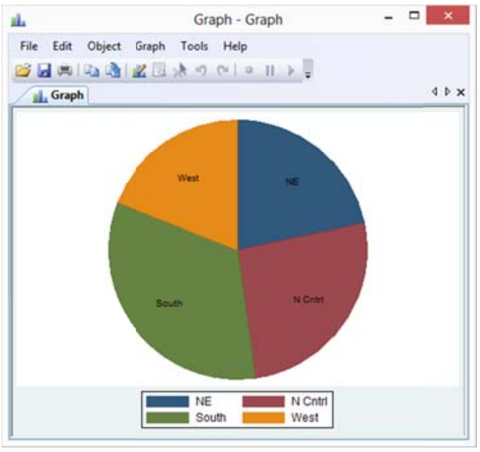
Examples include:

Scatter plots	Line plots
	

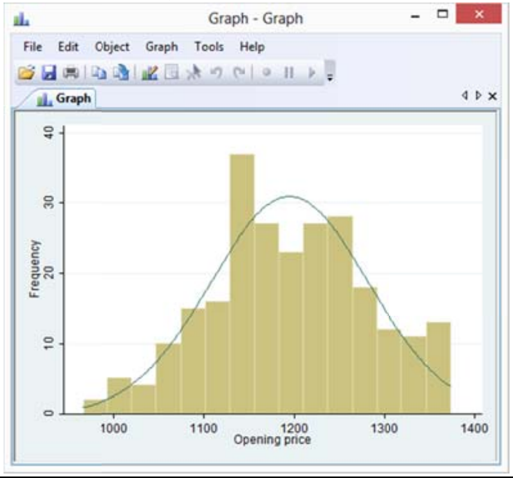
Bar charts



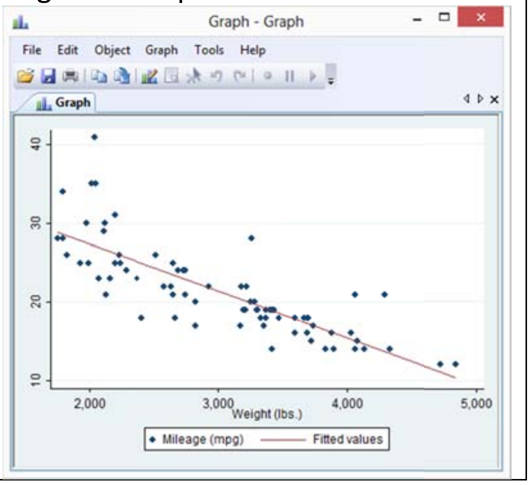
Pie charts



Distribution plots



Regression it plots



To begin wit we want to look at some distributional features of the variables in our data. For now let us concentrate on the age distribution of the children in our dataset.

1) What kind of graphs could be useful to look at the distribution?

2) Produce a histogram entitled ‘Age Distribution (DHS 2000 & 2005)’. The x-axis should display the age (in months) and be labelled accordingly.

3) You do have a strong colour preference and what the bars to be displayed in red.

4) Save the graph.

Child stunting is a major issue in Malawi. We therefore now want to graphically explore how serious this problem is for children, at different ages, from different backgrounds and how it evolved over time.

5) Investigate if there are any potential differences in stunting for boys and girls

6) Investigate the relationship between stunting and age.

7) Do children from wealthier backgrounds fare better? Is there a considerable difference?

8) How has stunting evolved over time?

Analysis with Stata

Following the graphical exploration in the following we want to identify and analyse the factors that influence stunting in children. In order to identify the determinants of stunting we will use standard OLS regressions.

For now we concentrate on the observations from 2000

1) In your data set which variables do you think have an effect on stunting?
   a. Group them in appropriate categories (i.e. Household characteristics, Location etc…).
   b. Identify the variables which are most useful.

2) How do you interpret your regression results?

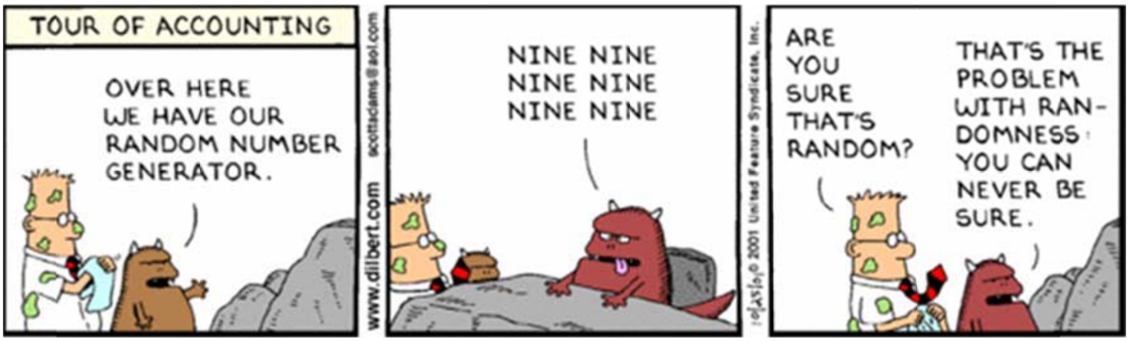
3) Are the same determinants also relevant in 2005? Any changes? How do you interpret those?

4) Present you regression output in a table.

Finally we are approaching the main objective of our analysis. Can you identify an effect of the food crisis?

Randomized selection methods

Test if you got the gist of the lecture going through the multiple choice questions at the end of Chapter 3 in *Kandhker* et al. (2010).


TOUR OF ACCOUNTING
OVER HERE WE HAVE OUR RANDOM NUMBER GENERATOR.
NINE NINE NINE NINE NINE NINE
ARE YOU SURE THAT'S RANDOM?
THAT'S THE PROBLEM WITH RANDOMNESS: YOU CAN NEVER BE SURE.
www.dilbert.com
scottadams@aol.com
© 2001 United Feature Syndicate, Inc.

1. Random Assignment

You are already familiar with the child data from the Malawi DHS in 2000. In the following we will use this information and randomly assign the children in the dataset into a treatment and control group. To do so follow the steps below and don't forget – a log and/or do-file might be handy ;).

1. Load the Malawi-data into Stata (use Tutorial 3 File which already contains all the dummy variables generated)

2. The variable 'caseid' provides a unique identification code for the mother of the children. The 'caseid' is constructed using the information from the variables v001 (cluster number), v002 (household number), and v003 (respondent line number). The caseid however is not unique for each child. You can see this if you sort the variable 'caseid'.

   sort caseid

   In your data editor you see that observations 2, 3 and 4 have the same caseid, for example. However, we want a unique identification number for each child which should take into account the birth order of the child (midx).

   Generate a new variable which combines the information for the 3 variables that make up the **caseid** and the **midx** into a unique ID **number** (not a string variable but a number) for each child. The new variable should be called 'ID'.

Hint: the number that you have to generate is quite long and the standard setting in Stata is that it only generates numbers that are 9 letters long. Hence, before you generate your variable you have to tell Stata that the number will be longer than 9 letters.
To do so start with the following sets of commands:

gen ID=. recast long ID format %13.0g ID

(Thereafter you can tell Stata how you want the ID to be constructed)

To check that all observations have a unique ID type:

duplicates list ID

If you have done it correctly, Stata should give you the following response (in the results window):

Duplicates in terms of ID
(0 observations are duplicates)

4. You now want to randomize, i.e. you want to randomly assign a number to each observation.

Seeding
If you want your analysis to be repeatable, you need to "seed" the random number generation process. This will result in exactly the same sequence of random number generated (e.g. 12345 or 28031986). Skip this step if repeatability is not necessary.

set seed 28031986 (see, if you can figure out who's date of birth that is)

Randomisation
To assign a random number to each individual observation, generate a new variable.

gen randomnbr = runiform() (the "runiform" randomly generates a number for each observation)

sort randomnbr

gen groupID = group(2) (This splits the sample into two groups, either assigned a 1 or a 2)

Alternative commands:

```
generate randomno = uniform()
egen grp2 = cut(randomno), group(2)
```

5. Now that you have randomly assigned each observation to a group, generate a dummy variable named “d_treatment” which takes the value 1 if the observation is in group 1 and will receive a treatment and 0 if the observation has been assigned to group 2 and will not receive the treatment.

2. Assessment

Your children have now been randomly assigned to a treatment (d_treatment=1) and a control group (d_treatment=0). If the randomisation worked well, both groups should on average be similar in their characteristics. Verify if that is the case. Do you note any differences? If so where and how would you interpret these?

If both groups are not the same and you will be asked to assess the impact of the programme how would you deal with the differences?

Randomized assessment

1. Assessment

The file Progressa.dta includes data from over 16,000 children in Mexico. The dataset includes information on both, children that benefited from Progressa and those that did not. You know that the implementation of Progressa was randomized.

- Check the balancing properties of your sample
- Assess the impact of Progressa on school enrolment of the children by:
  a. Comparing means

  b. Multivariate regression
     - Write down the equation you would estimate

  - Would you use an ordinary least squares model or is there a better alternative?
  - How would you do implement the estimation in STATA? What commands would you use?

- Looking at the results obtained, is the measured impact different between both methods, and if so, why do you think this is the case?

- What kind of effect have you actually estimated using the different approaches above?
  - ATE
  - ATT
  - ITE

  What is the difference between them?

- Assume that he participation in Progressa was not randomized. Instead, the programme is freely accessible to all eligible households. Nevertheless, programme information in form of leaflets has been randomly distributed. How could you assess the impact?

Regression discontinuity (RD)

In this session we will try to replicate the results from the paper “Do Voters affect or elect policies? Evidence from the U.S. House” by D. Lee, E. Moreti and M. Butler (2004, published in the Quarterly Journal of Economics).

The data to the paper is called LMB_data.dta.

In the paper they are not evaluating a policy intervention using an RD design but they use the method to examine the role of elections. The paper is not such an easy read so here just very briefly the intuition of the paper.

Starting point is that there are two fundamentally different views on the role of elections. Convergence assumes that due to the competition for votes, candidates seek compromise or middle ground policies (median voter theorem). The alternative, divergence, assumes that voters select candidates, who then enact their own preferred policies. So under convergence, **voters affect the policy choice** of the politicians whereas under divergence, **voters elect policies.** Which of the two we see in practise depends on whether politicians can make credible promises to implement policies which are not at their own bliss point (=Nutzenmaximum) (credible commitments are facilitated by repeated interactions). The goal of the paper is to examine, which of the two alternative views is more relevant for US politics, i.e. the voting in the House of Representatives.

Briefly to the theory (in simple terms, well at least as simple as possible):

- Assume we have 2 parties, D and R
  - R’s bliss point is 0, D’s bliss point is c(>0)
- The probability that D wins the election is P
- If D wins the election, policy x is implemented; if R wins, y is implemented
- P* represents the underlying popularity of party D, or the probability that D would win if x=c and y=0. An increase in P* represents an exogenous increase in D’s popularity
- When dx*/dP* and dy*/dP* > 0, we say that voters affect candidate’s policy choices
  - * denoted equilibrium
- When dx*/dP* and dy*/dP* = 0, we say that voters merely elect politicians with fixed policies. That is, an increase in P does nothing to the equilibrium policies of the parties.

The estimation Framework:

- The roll-call voting record (=namentliche Abstimmung) of the representative in the district following election t is
  - $RC_t = (1 – D_t)y_t + D_tx_t$
  - Where D_t is the indicator for whether D won. That is, only the winning candidate’s policy is observable
- The expression can be transformed into:

$$(2) \qquad RC_t = \text{constant} + \pi_0 P_t^* + \pi_1 D_t + \varepsilon_t$$

$$(3) \qquad RC_{t+1} = \text{constant} + \pi_0 P_{t+1}^* + \pi_1 D_{t+1} + \varepsilon_{t+1},$$

- This simply parameterizes the derivatives from the theory as π_0.
- It also allows an independent effect of party, π_1.

$_t$ and ε_t.

$$+ \underline{\pi_1[P^D_{t+1} - P^R_{t+1}]} = \gamma$$

$$(5) \qquad \underline{E[RC_t|D_t = 1] - E[RC_t|D_t = 0]} = \pi_1$$

$$(6) \qquad \underline{E[D_{t+1}|D_t = 1] - E[D_{t+1}|D_t = 0]} = P^D_{t+1} - P^R_{t+1},$$

- Everything underlined in red can be estimated from the data. Why this works?
  - The “elect component is $\pi_1[P^D_{t+1} - P^R_{t+1}]$
  - π_1 is estimated by the difference in voting records between the parties at time t
  - The fraction of districts won by Democrats in t+1 is an estimate of $[P^D_{t+1} - P^R_{t+1}]$
  - Because we can estimate the total effect, of a Dem victory in t on RC_{t+1}, we can then net out the elect component to implicitly get the affect component
  - Random assignment of D_t is crucial. Without it, equation (5) would reflect π1 and that Dem districts have more liberal bliss points

As an example on how to implement a regression discontinuity design we will try to replicate the results presented in Table I in their paper

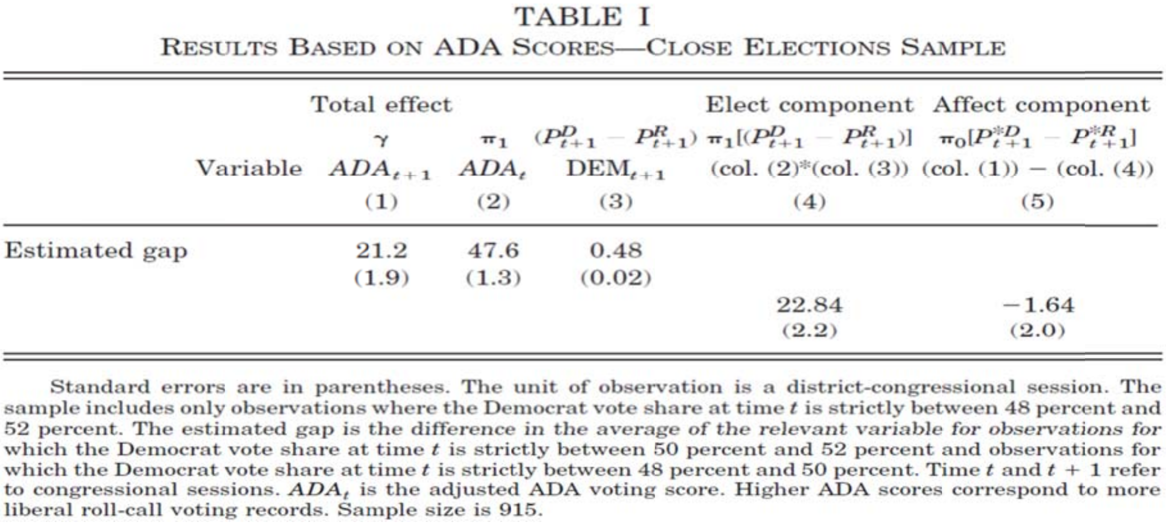


TABLE I
RESULTS BASED ON ADA SCORES—CLOSE ELECTIONS SAMPLE

	Total effect			Elect component	Affect component
	γ	π_1	$(P^D_{t+1} - P^R_{t+1})$	$\pi_1[(P^D_{t+1} - P^R_{t+1})]$	$\pi_0[P^{*D}_{t+1} - P^{*R}_{t+1}]$
Variable	ADA_{t+1}	ADA_t	DEM_{t+1}	(col. (2)*(col. (3))	(col. (1)) − (col. (4))
	(1)	(2)	(3)	(4)	(5)
Estimated gap	21.2 (1.9)	47.6 (1.3)	0.48 (0.02)		
				22.84 (2.2)	−1.64 (2.0)

Standard errors are in parentheses. The unit of observation is a district-congressional session. The sample includes only observations where the Democrat vote share at time *t* is strictly between 48 percent and 52 percent. The estimated gap is the difference in the average of the relevant variable for observations for which the Democrat vote share at time *t* is strictly between 50 percent and 52 percent and observations for which the Democrat vote share at time *t* is strictly between 48 percent and 50 percent. Time *t* and *t* + 1 refer to congressional sessions. ADA_t is the adjusted ADA voting score. Higher ADA scores correspond to more liberal roll-call voting records. Sample size is 915.

And the graph II.A

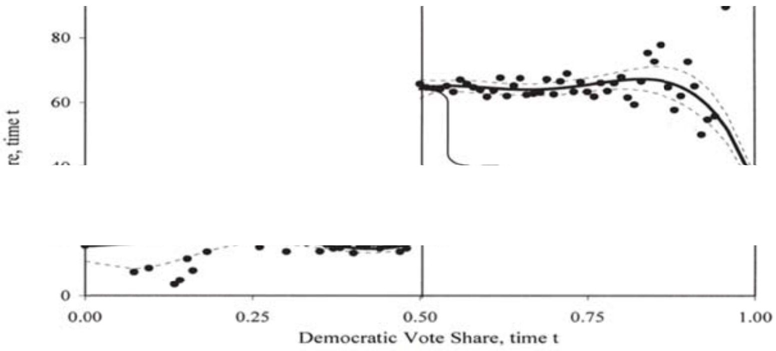


Figure IIa
Effect of Party Affiliation: π_1

<u>Task:</u>

- To begin install the packages **rd** and **cmogram**. Use the findit function in STATA.
- Use the help command to get a description of the **rd** and **cmogram** commands to see how they are structured.
- The outcome variable you want to evaluate from your data is score (this is the ADA score that the authors use in the paper), what would your STATA command look like in order to get an impact estimate assuming you looking at values close to the cut-off of 0.5 of the democratic vote share (the variable in the data is called demvoteshare), i.e. with a bandwidths between 0.48-0.52.

For further reading a good paper on RD designs, their implementation and shortcomings is:

David S. Lee & Thomas Lemieux (2010). "Regression Discontinuity Designs in Economics," Journal of Economic Literature, 48(2): 281-355

Difference-in-Difference (DID)

Difference-in-Difference with cross-section data:

The Malawi.dta contains information on children (0-5 years) surveyed in the Demographic and Health survey in 2000 and 2004/5. In each year the children to be surveyed where randomly selected. Thus, in the data you do not observe the same children twice but you have a different group of children at each point in time. Hence, the Malawi.dta contains cross-section information or given that it combines two sets of cross-section data we also call it pooled cross-section data.

In 2002 Malawi experienced a major food crisis due to crop failure leaving about 2.5 million people at the verge of starvation. Past research shows that particularly the first 2-3 years in life are crucial for later development and that negative health shocks encountered during these early years in childhood can have lasting consequences for outcomes later in live. Against this background you are interested in investigating the impact of the food crisis on child health to see if there are any long-term negative consequences to be expected. The outcome you are interested in is the weight-for-height z-score. The Weight-for-height z-score (WHZ) is a measure of the deficit in tissue and fat mass and is sensitive to temporary food shortages and episodes of illness and therefore a frequently used indicator to identify short-term malnutrition. Usually children with WHZ below -2SD are considered to be malnourished.

Given that not all regions of have been hit equally bad by the crisis at the time and some regions have even been spared we can use this variation in the severity of the shock to construct a difference-in-difference estimator to identify the effects of the food crisis.

The basic estimation equation to be implemented is as follows:

$$y_i = \beta_0 + \beta_1 C_i + \beta_2 M_i + \beta_3 H_i + \eta_n + \delta_1 dT_n + \delta_2 dS_t + \delta_3 dS_t dT_n + u_i,$$

where y represents the outcome of interest, weight-for-height (WHZ). The subscript i indicates that the outcome varies per individual, n by district and t by time period. With T representing the treatment group, the dummy dT captures possible permanent differences between the affected and non-affected regions, the dummy dS indicates the survey year (2000=0, 2004=1) and thus is a time-effect absorbing aggregate factor that would cause changes in the outcome variable over time for all observation units, even in the absence of a shock or intervention. The effect of interest is the so-called treatment effect, δ_3, which is associated with the interaction between belonging to the treated group and the time effect. More precisely, it captures the difference in means within the treatment and control group before and after the crisis and thus provides an estimate of the impact of the food shortage on the respective outcome variable.

In order to increase the precision of the estimate, a number of control variables have been included in the specification, accounting for potential observable differences between the treatment and control group. C is a vector of child characteristics e.g. gender of the child, age etc. M includes information about the mothers, i.e. the years of education. The vector H contains household specific information, i.e. the wealth category, or whether the head of the household is a woman.

- The variable d_2004 is a dummy variable indicating whether the child has been survey in 2000 (=0) or 2004 (=1). The variable d_affected indicates whether the child was living in a region that was hit by the crisis or not. In order to implement the above estimation equation in Stata you have to generate a new variable representing the interaction term. To do that type:

  *gen interaction=d_2004*d_affected*

- Now you can implement the estimation equation using the *reg* command as you have already seem several times now.

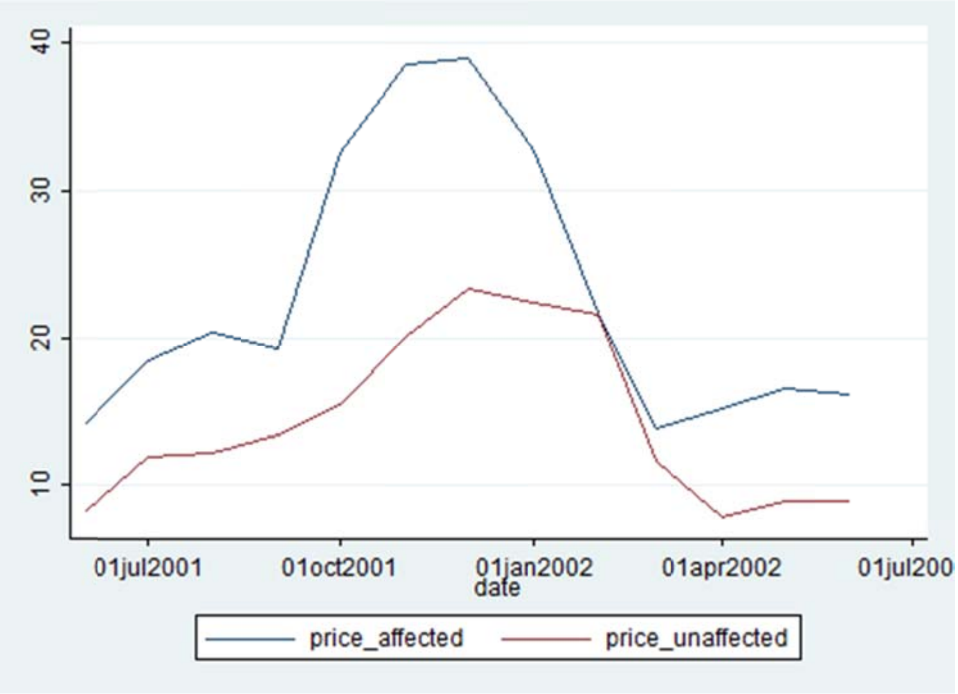
  reg whz male childage_month …. d_2004 d_affected interaction, robust cluster(v001)

- The coefficient on the interaction gives you the impact estimate. How would you interpret the results you got?

- Given that children of different ages or male and female children might have been differently affected by the crisis, try to estimate the effect of the food crisis separately for male and female children and children aged between 6-24 months and above 24 months separately.

For a difference-in-difference estimation to give you an unbiased estimate, one assumption that has to hold is the so called parallel trends assumption. This basically means that prior to the event or intervention, the treatment and control groups should have developed similarly over time. In our example, we the affected areas are those in which maize prices between January and March 2002 were particularly high. The data set Prices.dta contains information on the average prices in the affected and unaffected regions.

- Using this data plot a graph which shows the development of the Maize price in the regions over time. Have a go at it using the graphics menu you saw in the last tutorial. You should get a graph looking like this:



Based on the graph, would you say that the parallel trends assumption holds in our case?

Difference-in-Difference with panel data:

In this case we are evaluating the impact of microcredit in rural Sri Lanka. The dataset SriLanka_9198.dta contains data on 826 households which have been repeatedly sampled in 1991 and 1998. The variable "credit" indicates whether the household has benefited from a microcredit by 1998 or not.

Your task is to evaluate the impact of the microcredit on household food consumption.

1. Have a look at your data and the information you have. Identify which variables will be important in order to apply a difference-in-difference approach analogue to the cross-section example above.

2. Write down the regression function which you would use to estimate the effect of microcredit on food consumption. Which control variables would you include and why?

3. Start by running a basic regression with the outcome variable and the difference in difference operators. Your command should be:

   reg expfd year credit interaction, robust

   - Which coefficient gives you the impact estimate?

- How do you interpret the regression results?

4. Often, when looking at expenditure as outcome you will see that researchers are not using the absolute values but that they are using log(expenditure).

   - Why would you do that?
   - Generate a new outcome variable called lexpfd which is the natural log of the expenditure variable.

     gen lexpdf=ln(expdf)

   - Rerun the basic regression from above with log-food expenditure as outcome. How does this change the results? How would you interpret the regression results now?

5. Now, include control variables in your regression. How does this change your impact estimate?

6. The microcredit programme has not been implemented using random assignment, instead households are free to chose whether to ask for a loan or not. Given that we have no information on the underlying motivation on why some households asked for credit and others did not, the results that you have obtained so far night be influenced by something called "selection bias".

   In the panel data that you have you do observe the same households at two points in time. This situation allows to control for factors that are actually unobserved but which remain constant overtime. (Basically by subtracting the actual condition of the household from the initial condition all factors that remained stable at the two periods should be netted out). This is done by using a fixed effects model. The basic command would be:

   xtreg lexpfd year credit interaction, fe i(nh)

   xtreg is a panel data command. Have a look at the description of this command using the help function in Stata. The "fe" indicates that you are running a fixed effects model. With the "i(nh)" you tell Stata which variable identifies the household. Remember, the household ID has to be unique for each household interviewed in 1991 and 1998 respectively. Do you remember how to check for duplicates?

7. Now include control variables in your fixed effects regression. Are the results from the fixed-effects regression different to the ones you got before from the pooled regression?

Propensity score matching (PSM)

The basic idea behind propensity score matching (PSM) is to match each participant with an identical non-participant and then measure the average difference in the outcome between the participant and the non-participant.

As example we work with data from Bangladesh also used by Khandker et al. (2010). The data (hh_98.dta) is used to assess the impact of micro-finance. The variable which determines programme participation is "dmmfd" i.e. indicating whether the male household member has received a micro-credit or not.

The first step in implementing a propensity score matching approach is to determine the propensity score.

There are different ways this can be done.

1. Using a probit/logit model

Basically the propensity score is nothing else then the predicted value of a regression of programme participation (dependent variable) on a set of pre-programme outcomes. You already know how to run a logit/probit regression in Stata so which control variables would you consider?

To get the predicted value type

Predict p

Stata then automatically generates a new variable called "p". You can change the name and call it propensity_score.

2. Using the pscore command

Another alternative is to use the pscore command. This is a user-writen command. Check if it is already installed in your Stata version.

It works similar to the regression command i.e. after pscore type the dependent and the independent variables.

pscore dmmfd sexhead agehead educhead vaccess pcirr rice wheat milk oil egg [pw=weight], pscore(ps98) blockid(blockf1) comsup level(0.001)

After the comma, the values in brackets are just the names of new variables generated. So here we call the propensity score ps98. Comsup identifies the region of common support.

The advantage of the pscore command is that automatically checks if the balancing property is satisfied, i.e. if your treatment and control household are on average the same.

Check the regression output and see if your balancing property is satisfied. If not what do you do?

- Compare the propensity scores obtained from pscore and the probit regression.

3. Using the psmatch2 command

A third alternative would be the *psmatch2* command this one allows you to estimate the propensity score but at the same time also already do the matching. This is also a package you have to install in Stata separately. You can use the *findit* function and follow the instructions on screen or use *scc install* like you saw in Tutorial 5.

Nearest Neighbour matching:

When doing nearest neighbour matching we compare participants to non-participants with the closest propensity score. Let’s estimate the average treatment effect of credit on landholding.

```
psmatch2 dmmfd, out(hhland) pscore(ps98) common
```

An alternative specification would be:

```
psmatch2 dmmfd, out(hhland) pscore(ps98) n(5) common
```

Here we are comparing the participant to its 5 nearest neighbours.

Kernel:

Apart from nearest neighbour matching, a common matching algorithm used is kernel. This basically applies a kernel function and weighting to “match”.

```
psmatch2 dmmfd, kernel out(hhland) k(tricube) pscore(ps98) common
```

Checking the robustness of your matching results

There are several ways to check the robustness of your findings. One approach is to use different matching methods as you have just done. So if you compare your results are they robust?

Further excercise:

- Now that you have assessed male microcredit participation, assess the effects of female microcredit participation on expenditure.

- Think about last weeks exercise on the difference-in-difference estimation. In order to improve your results or check the robustness you could implement a difference-in-difference estimation with matching. How would you do it?

Instrumental variable (IV)

IV is just another method that can be used for impact evaluation when treatment is not randomly assigned. The tricky part about using IV is to find an appropriate instrument. So, you want to find an observable variable or variables that are exogenous, i.e. that influence the participation in the programme you want to evaluate but that do not influence the outcome itself.

Take the example from last time again (hh_98.dta). Now that you are all pros in Stata ☺, how would you go about estimating the impact of the microcredit programme using an IV approach?

- What would be a suitable instrument(s)? Remember it has to be exogenous.

- Write down your econometric equation.

- You know that IV-estimation is a two-step process. You can implement it using the standard regression commands you have already seen. How would you do it?

- Of course, as always, there is also already a command that allows you to implement IV or rather 2SLS regressions straight away in one go – *ivreg*. Use this one to estimate your regression.

- Now, to the most important thing of course, the IV regression results stand and fall with your instrument. How can you test and ensure that you have a good instrument?

A bit of this and that…

1. Sample size calculations in a randomized design

Some key terms to remember:

1. Power: the likelihood that, when the program has an effect, one will be able to distinguish the effect from zero given the sample size.
2. Significance: the likelihood that the measured effect did not occur by chance. Statistical tests are performed to determine whether one group (e.g. the experimental group) is different from another group (e.g. comparison group) on the measurable outcome variables used in the evaluation.
3. Standard Deviation: a standardized measure of the variation of a sample population from its mean on a given characteristic/outcome. Mathematically, the square root of the variance.
4. Standardized Effect Size: a standardized measure of the [expected] magnitude of the effect of a program.
5. Cluster: the level of observation at which a sample size is measured. Generally, observations which are highly correlated with each other should be clustered and the sample size should be measured at this clustered level.
6. Intra-cluster Correlation Coefficient: a measure of the correlation between observations within a cluster; i.e. the level of correlation in drinking water source for individuals in a household.

An example:
In India they have been introducing a scheme of tutoring in school whereby they send tutors to school to attend classes and help the weakest children in class to understand the course material etc. These tutors are called Balsakhis. We are interested in measuring the impact of the Balsakhis in class on children's school outcomes, i.e. test scores measured at the individual level. Balsakhis were randomly assigned to a class, i.e. the randomization was done at the classroom level. It could be that our outcome of interest is correlated for students in the same classroom, for reasons that have nothing to do with the Balsakhi. For example, all the students in a classroom will be affected by their original teacher, by whether their classroom is unusually dark, or if they have a chalkboard; these factors mean that when one student in the class does particularly well for this reason, all the students in that classroom probably also do better—which might have nothing to do with a Balsakhi.

Therefore, if we sample 100 kids from 10 randomly selected schools, that sample is less representative of the population of schools in the city than if we selected 100 random kids from the whole population of schools, and therefore absorbs less variance. In effect, we have a smaller sample size than we think. This will lead to more noise in our sample, and hence larger standard error than in the usual case of independent sampling. When planning both the sample size and the best way to sample classrooms, we need to take this into account.

This exercise will help you understand how to do that. Should you sample every student in just a few schools? Should you sample a few students from many schools? How do you decide?

We will work through these questions by determining the sample size that allows us to detect a specific effect with at least 80% power. Remember power is the likelihood that when the treatment has an effect you will be able to distinguish it from zero in your sample.
In this example, “clusters” refer to “clusters of children”—in other words, “classrooms” or “schools”. This exercise shows you how the power of your sample changes with the number of clusters, the size of the clusters, the size of the treatment effect and the Intraclass Correlation Coefficient.

Using the OD Software:
Download the OD software from the website (a software manual is also available):

http://sitemaker.umich.edu/group-based/optimal_design_software

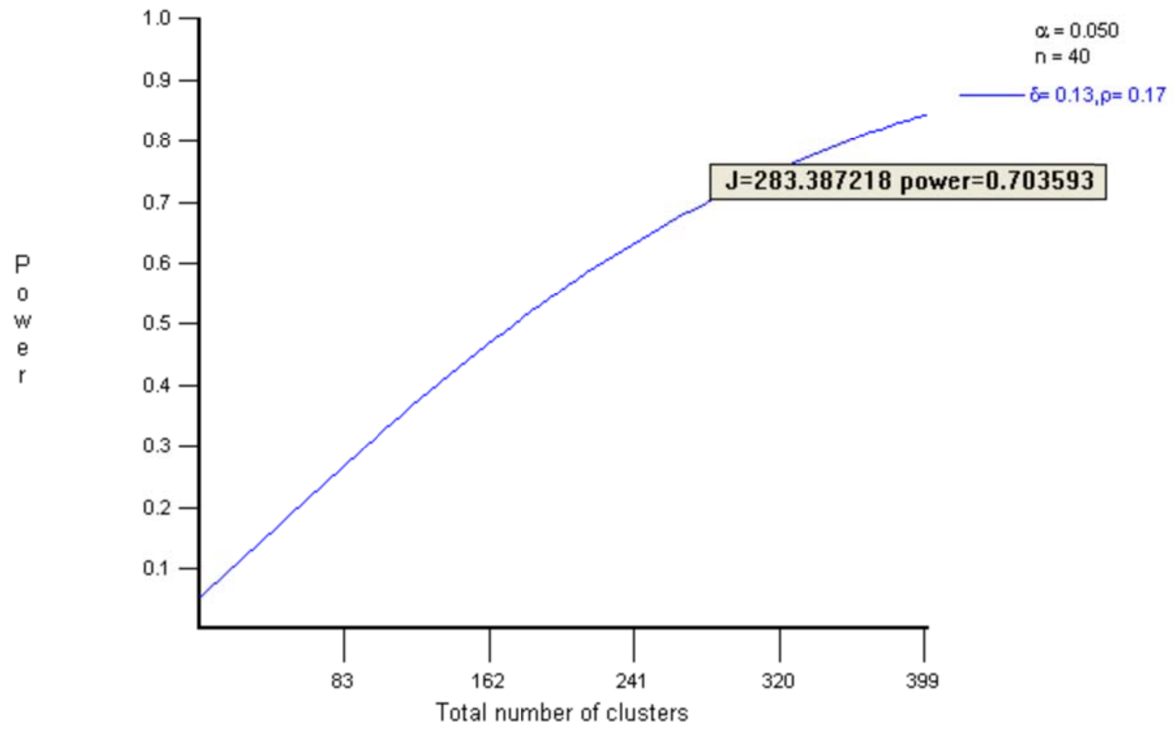
When you open it, you will get a window with a blank screen. On top select the menu option “Design” to see the primary menu. Select the option “Cluster Randomized Trials with person-level outcomes,” “Cluster Randomized Trials,” and then “Treatment at level 2.” You’ll see several options to generate graphs; choose “Power vs. total number of clusters (J).”

A new window will appear. Select α (alpha). You’ll see it is already set to 0.050 for a 95% significance level.

First let’s assume we want to test only 40 students per school. How many schools do you need to go to in order to have a statistically significant answer? Click on n, which represents the number of students per school. Since we are testing only 40 students per school, so fill in n(1) with 40 and click OK.

Now we have to determine δ (delta), the standard effect size (the effect size divided by the standard deviation of the variable of interest). Assume we are interested in detecting whether there is an increase of 10% in test scores. (Or more accurately, are uninterested in a detect less than 10%) Our baseline survey indicated that the average test score is 26, with a standard deviation of 20. We want to detect an effect size of 10% of 26, which is 2.6. We divide 2.6 by the standard deviation to get δ equal to 2.6/20, or 0.13. Select δ from the menu. In the dialogue box that appears there is a prefilled value of 0.200 for delta(1). Change the value to 0.13, and change the value of delta(2) to empty. Select OK.

Finally, we need to choose ρ (rho), which is the intra-cluster correlation. Ρ tells us how strongly the outcomes are correlated for units within the same cluster. If students from the same school were clones (no variation) and all scored the same on the test, then ρ would equal 1. If, on the other hand, students from the same schools are in fact independent—and there were no differences between schools, then ρ will equal 0. You have determined in your pilot study that ρ is 0.17. Fill in rho(1) to 0.17, and set rho(2) to be empty.



You'll notice that your x axis isn't long enough to allow you to see what number of clusters would give you 80% power. Click on the sixth button from the left to set your x axis maximum to 400. Then, you can click on the graph with your mouse to see the exact power and number of clusters for a particular point.

Now you have seen how many clusters you need for 80% power, sampling 40 students per school. Suppose instead that you only have the ability to go to 124 schools (this is the actual number that was sampled in the Balsakhi programme).

Finally, let's see how the Intraclass Correlation Coefficient (ρ) changes power of a given sample. Leave rho(1) to be 0.17 but for comparison change rho(2) to 0.0.

To take a look at some of the other menu options, close the graph by clicking on the **X** in the top right hand corner of the inner window. Select the Cluster Randomized Trial menu again…

2. Another handy tool in Stata

- Instead of copy and pasting your regression tables you can directly transfer them into word or tex (that latter one is when you are really approaching geek status). For that you need to install a package called "outreg"
  Run a regression with any of the data you have at hand e.g. hh_98.dta

outreg using "put the name of the file" (as options, replace, append, etc.)

Any questions/wishes?

Case Study: Getting Parents Involved

This case study is based on "Getting Parents Involved: A field Experiment in Deprived Schools" by Francesco Avvisati, Marc Gurgand, Nina Guyon, and Eric Maurin (CEPR Discussion Paper, 8020, 2010).

The material comes from the J-PAL course "Evaluating Social Programmes".

Key Vocabulary

1. **Hypothesis**: a proposed explanation of and for the effects of a given intervention. Hypothesis are intended to be made ex-ante, or prior to the implementation of the intervention.
2. **Indicators**: metrics used to quantify and measure specific short-term and long-term effects of a programme.
3. **Logical Framework**: a management tool used to facilitate the design, execution, and evaluation of an implementation. It involves identifying strategic elements (inputs, outputs, outcomes and impact) and their causal relationships, indicators, and the assumptions and risks that may influence success and failure.
4. **Theory of Change**: describes a strategy or blueprint of achieving a given long-term goal. It identifies the preconditions, pathways and interventions necessary for an initiatives success.

Background

Problems of truancy and discipline can contribute to many school children in industrialised societies graduating from school without mastering basic skills. The school district of Creteil, in France, is a densely populated area with high proportions of immigrants from mostly Maghreb countries, and has very poor socioeconomic indicators. In such a setting, linguistic and social barriers along with financial and logistical constraints can prevent parents from paying closer attention to their children's education.

Increasing parental involvement has been widely touted as a means of overcoming difficulties in child learning behaviour. The programme called "*La mallette des parents*", was designed to foster parental involvement through a series of monthly meetings with the school staff on how to successfully manage the transition from primary to middle school. These discussions provided parents of sixth grades with information on the French school system and guidelines on how to assist children with homework.

The experiment should answer the following questions:

- Can parental involvement be used as a lever to improve educational outcomes in France?
- Does greater engagement by parents improve discipline and behaviour?
- Do classroom interactions also result in positive effects for children whose parents don't attend the meetings?

The French Educational Environment

The French state-run educational system is highly centralised with schools having limited autonomy. All schools teach the same education curriculum and employ teachers that are selected through national examinations. There is no tracking of students by ability and French parents are not free to choose what school their children will attend.

The pool of students in the district of Creteil, where the programme took place, is very heterogeneous in ability and diverse in cultural backgrounds. These suburbs East of Paris have large populations of recent and second-generation immigrants. A recent survey showed that more than 20% of the local population is first-generation immigrants, and many are relatively poor. These parents face numerous barriers to navigating the hierarchical education system: many speak limited French and work far away from local schools. This lack of parental involvement might be the cause of problems like truancy and indiscipline in children, especially in the poorer districts, where many pupils are still far from reaching the basic requirements of curricula (OECD, 2010).

Informational Campaign for Parents

Just after the start of the academic year, schools sent informational leaflets to families of sixth graders asking them to register for a series of meeting with school staff on how to successfully navigate the transition to middle school. (Not all schools were planning to offer this programme, however. The registration was less of an enrolment process and more of a tool to gauge the level of interest from parents.) Those schools that ultimately participated would offer a series of three meetings over the course of three months.

After registration, the families in participating schools were given an offer to continue with one of two additional programmes, or to abstain from further involvement:

- Programme A: An additional series of monthly meetings that complement the three initial meetings. Parent and school are encouraged to invite external experts to theses meetings.
- Programme B: An additional series of more intense meetings held as often as twice a week for four or five months. These meetings focus on providing training for parents needing further support to improve their literacy or computer skills.

The Scope of Discussions

The goal of these highly interactive meetings was to help parents understand the role of each member of the educational community, the school’s organisation, and to help them develop positive involvement attitudes towards their children’s education. Facilitators were given standard materials, including a DVD explaining the purpose of various school personnel and documents explaining the functions of the various school offices. The two initial sessions focused on how parents could help their children with homework and the last session took place after the distribution of report cards to help them adapt to the first term results and to give them tips on how to go forward.

Task

Your evaluation team has been entrusted with the responsibility of evaluating the campaign's impact on child learning and behaviour. Your evaluation should address all dimensions in which informational campaigns for parents can affect cognitive and non-cognitive abilities of children. How might the meetings encourage greater involvement by parents? What are the most important outcomes to test? What steps must occur for these changes to take place? What data should your team collect to evaluate the intervention?

Discussion Topic 1: Needs

1. Who is the target population?
2. What characteristics of the French educational system make it particularly challenging for the students?
3. What features of the home environment make it challenging?
4. What might we see different in households of high-performing students

Discussion Topic 2: Programme Theory

1. What are the main characteristics of the informational meetings?
2. How might these meetings encourage parents to pay more attention to their children's education?
3. What are the potential challenges? Why might the programme fail?

Discussion Topic 3: Outcomes and Indicators

1. What are the possible positive, negative and null effects of the intervention on child development and learning?
2. List all indicators that you would use to measure each of the potential outcomes?

Discussion Topic 4: Defining the Hypothesis

1. What might be some examples of key hypotheses you would test? Pick one.
2. Which indicators would you use to test your primary hypothesis?

Discussion Topic 5: Formalizing the Theory of Change

1. What are the steps and conditions that link the informational campaign for parents to the final outcomes?
2. What indicators should you measure at each of these steps?
3. Using the outcomes and conditions, draw a logical framework/results chain, linking the intervention to the final outcomes.