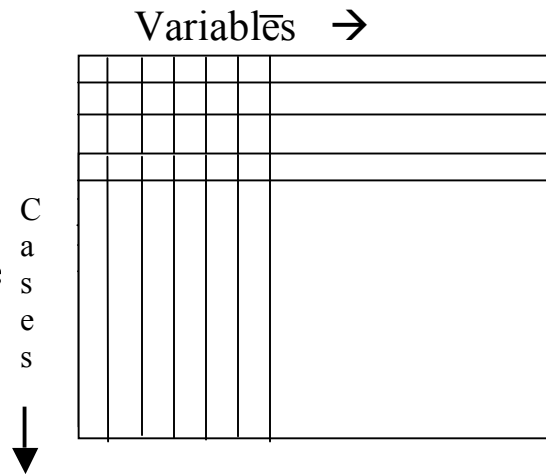


Vocabulary of data analysis

Data – data matrix – dataset

Cases – variables – codebook – system file



Let's see how this looks with STATA.

Double click on the file you downloaded from my website. If you have STATA installed on your computer, it will open, showing a screen that looks something like...

Intercooled Stata menus and toolbar	
Review	Results
Variables	
	Stata command

Things to do on setting up STATA

In Edit → Preferences → general preferences

Change Data Editor by unchecking "Copy value labels instead of numbers", click "Apply"

(Include variable names on copy to clipboard should remain checked)

In Edit → Preferences → Windowing

Set Results Window to 999999 (leave checkboxes unchecked for now). Click "Apply"

(Stata will reduce this large number to whatever maximum is allowed)

When asked, DO NOT click the line that tells STATA to go ahead and update its code/dofiles, but leave that window open so you can do it later when you have more time.

Type 'set memory 200M, permanent' -- note the comma, don't type the quote marks

Strategies of STATA usage: ALWAYS LOG YOUR SESSION

(but change the type of file to .log so we can use them in class)

The basic syntax of a STATA command:

[prefix:] command [varlist] [=expression] [**if...**] [**in...**] [weight] [**using...**], options

The easiest way to see how commands are constructed is to use the menus in order to create a few simple ones

Choose Statistics → summaries → summary and descriptive → summary statistics

Then click on the names of the variables you want summarized (I suggest ‘elect’ ‘turnout’ and ‘timeleft’ – these are all numeric variables for which a summary makes some sense)

Click ‘OK’ to have the command performed. You will see the text of the command printed in the Results window before the output from the command.

You can retrieve the command into your command line by pressing the ‘page up’ button on your keyboard.

Change the variable list, eg by double clicking on 'yrsleft' and replacing it with 'natturn'

The results window contains the mean and standard deviation of each variable, as well as the minimum, maximum, and the number of observations (cases) over which these statistics were calculated.

What is a mean? In statistical notation the mean is indicated by putting a bar across the top of the letter standing for the variable whose mean it is. So the mean of X is

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N}$$

Here, i stands for "individual" and the big Greek sigma stands for "sum" while N stands for "Number of individuals".

The i=1 and N at bottom and top of the sigma tells us we are summing across all individuals from 1 to N (the number of individuals).

In other words, the mean is just the average, but the notation is important. The letter sigma and how we use it represents one of about five pieces of mathematical notation that you will need to master.

Standard deviation. Consider the following table of numbers. Each column has the same mean, but the extent to which that mean accurately portrays the set of numbers in the column is different...

Individual	X	Y	Z
1	5	4	2
2	5	5	5
2	5	6	8
Mean =	5	5	5

With large Ns we need a measure that tells us how much the values are dispersed – the range is a measure of dispersion, but does not take account of how values are dispersed between the extremes.

The standard deviation is represented by the roman letter S, for Standard deviation, and is generally subscripted with the variable whose standard deviation we are talking about. So

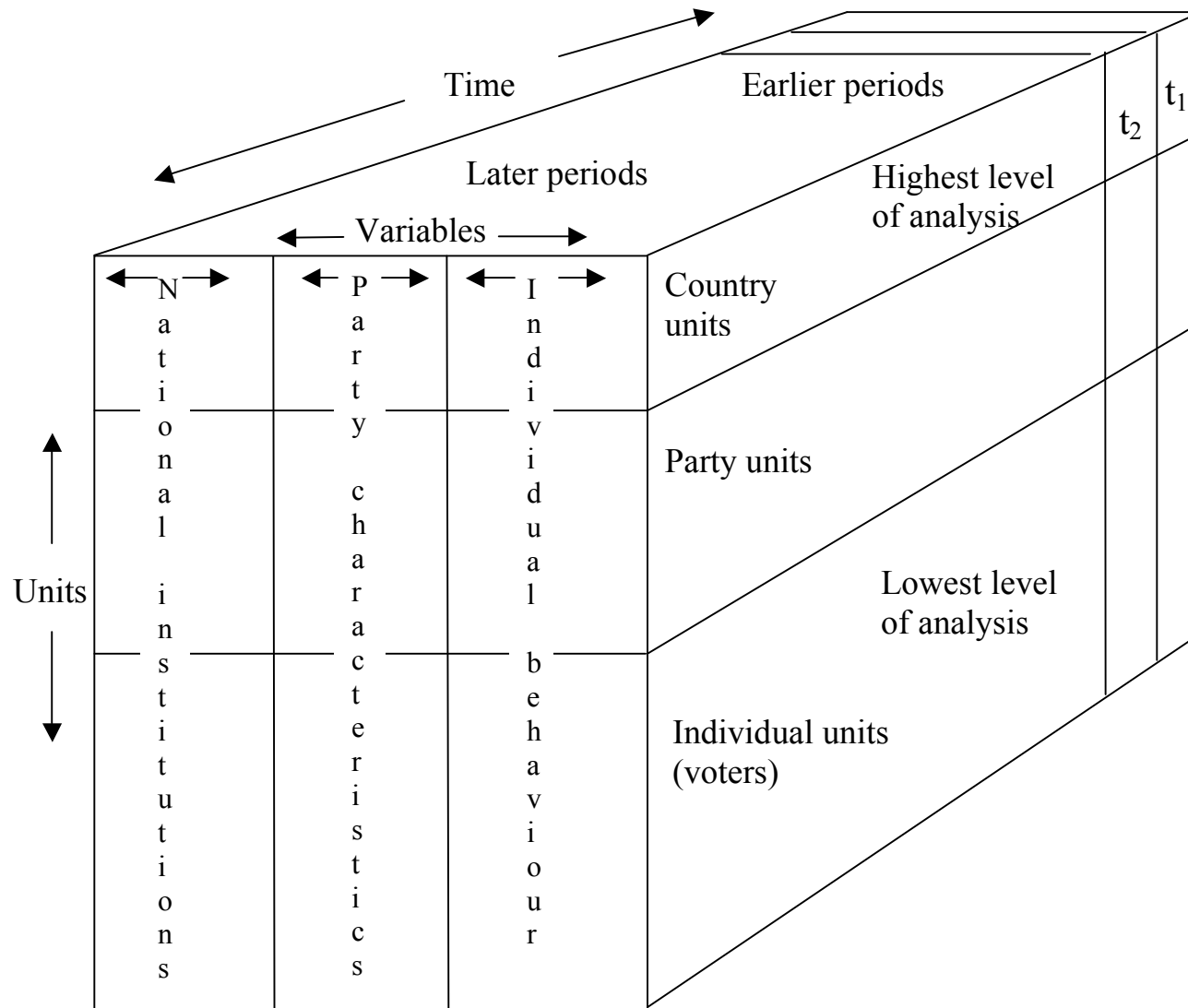
$$S_x = \sqrt{\frac{\sum (X_i - \bar{X}.)^2}{N}}$$

If you square S (or don't take the square root in the above) you get the **variance**, S^2 (see below).

Digression on Levels of Analysis:

individual – party – country (or individual – classroom – school)

The Complete Data Box



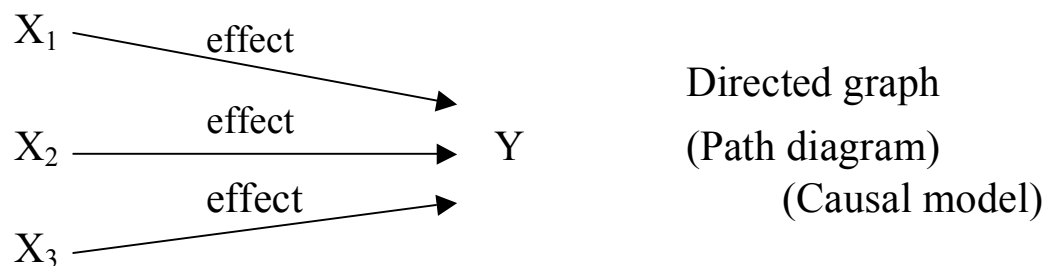
Levels of measurement

Interval – ordinal – nominal – dummy

Type of variable	Distinguishing characteristic
Interval	Unit of measurement (2 is \$1 more than 1)
Ordinal	Scale of measurement (2 is more than 1)
Nominal	Mutual exclusivity (2 rules out 1)
Dummy (Dichotomy)	Attribute (0=no, 1=yes)

Dependent and independent variables

Independent → Dependent



X_2 and X_3 might be viewed as 'control variables'

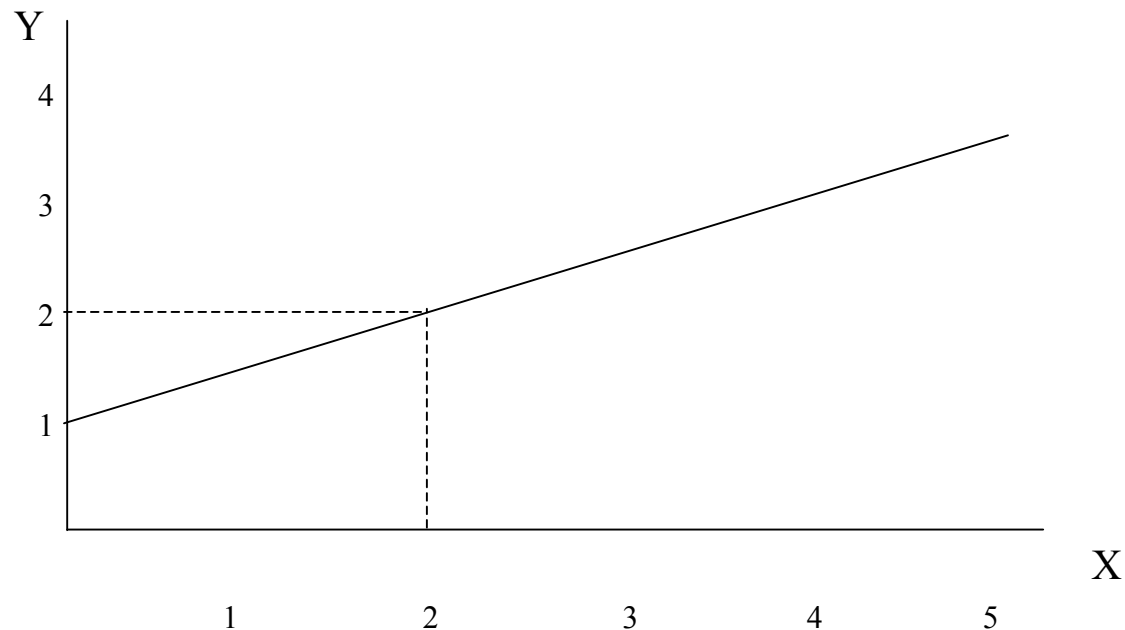
Each arrow can have a coefficient showing the strength of the effect carried by that path.

REGRESSION EQUATION:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$$

Prediction equation
(regression equation)

The regression coefficients (bs) are the effects in a path diagram .But a regression equation gives us the raw materials for more than a **directed graph**, it also gives us the raw material for a **Cartesian graph**



This is the Cartesian graph for $Y = 1 + 0.5 X$

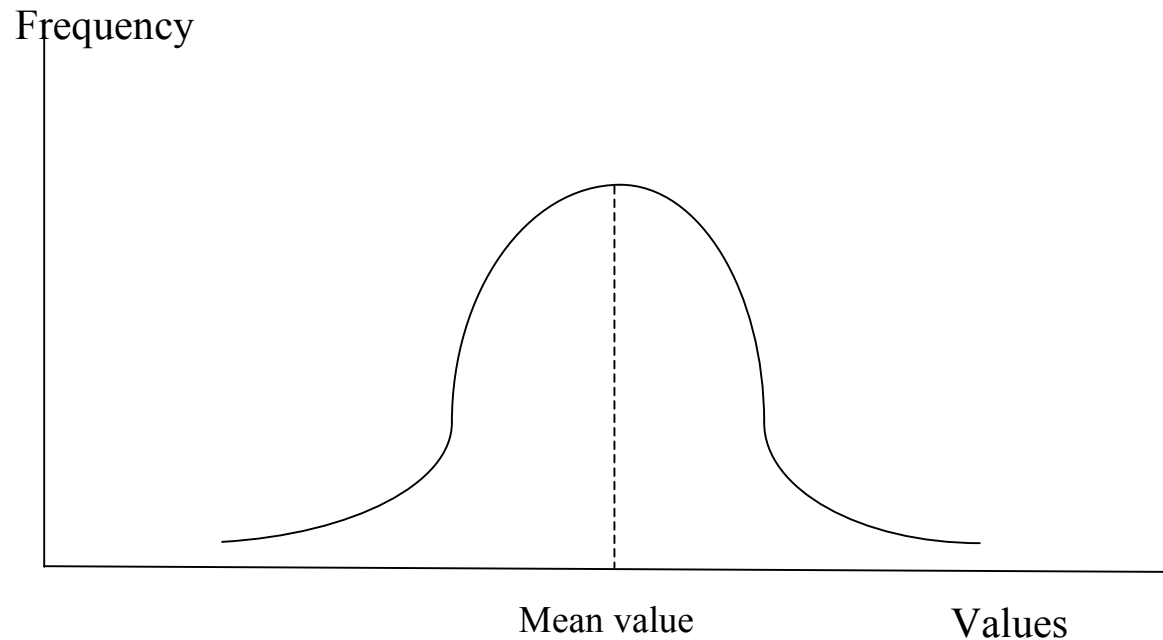
When a Cartesian graph is drawn on the basis of a regression analysis, the sloping line is called the “best-fitting line”

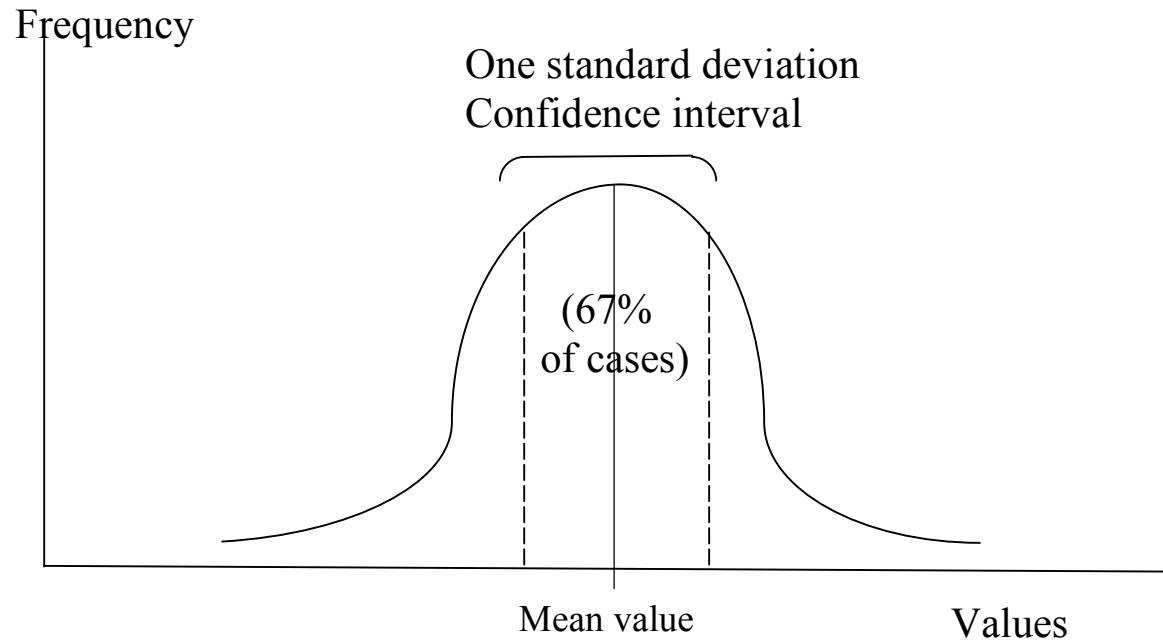
Variance explained (R^2) – the proportion of the variance in the dependent variable explained by the independent variable(s). The variance about the best-fitting line.

Significance level: The probability of being wrong when you conclude that there is a relationship between two variables (or an effect of one variable on another)

Confidence intervals: The span (+ or -) between which a coefficient falls with given probability.

Frequency distribution and normal curve (not drawn accurately)...





The "normal" curve is designed so that 66.66666% of cases fall within one standard deviation above or below the mean, 95% of cases fall within two standard deviations of the mean. The two-standard deviation confidence interval is the interval that corresponds with the industry standard of 95% confidence (or a 5% chance of being wrong).

Variance explained

This has to do with how much of the overall variance is accounted for by some relationship.

$$R^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

What is explained variance?

The variance that is left after the mean(s) has/ve been moved to account for some relationship

It is the square of the product moment correlation (r)

Size of coefficient*	Interpretation with Individual-level data	Interpretation with Aggregate data
$r/R/R^2 = 0.00$ to 0.06	Trivial	Trivial
$r/R/R^2 = 0.07$ to 0.19	Slight	Trivial
$r/R/R^2 = 0.20$ to 0.34	Moderate	Slight
$r/R/R^2 = 0.35$ to 0.49	Strong	Moderate
$r/R/R^2 = 0.50$ to 0.65	Spectacular	Strong
$r/R/R^2 = 0.66$ to 0.80	Highly spectacular	Very strong
$r/R/R^2 = 0.81$ to 0.95	Suspect	Spectacular
$r/R/R^2 = 0.96$ to 1.00	Very suspect	Very suspect

*Note: Interpretations apply to r/R for bivariate analysis, R^2 for multivariate analysis.

Strategies of STATA usage: ALWAYS LOG YOUR SESSION

(but always change the type of file to .log so we can use them in class)

The basic syntax of a STATA command:

[prefix:] [quietly] command [varlist] [=][expression] [**if...**] [**in...**] [weight] [**using...**], options

The easiest way to see how commands are constructed is to use the menus in order to create a few simple ones

Choose Statistics → summaries → summary and descriptive → summary statistics

Then click on the names of the variables you want summarized (I suggest ‘elect’ ‘turnout’ and ‘timeleft’ – these are all numeric variables for which a summary makes some sense)

Click ‘OK’ to have the command performed. You will see the text of the command printed in the Results window before the output from the command.

You can retrieve the command into your command line by pressing the ‘page up’ button on your keyboard.

Change the variable list, eg by double clicking on ‘yrsleft’ and replacing it with ‘natturn’

The results window contains the mean and standard deviation of each variable, as well as the minimum, maximum, and the number of observations (cases) over which these statistics were calculated.

Retrieve the command by typing the ‘page-up’ button on your keyboard, or click on the line in the review window.

Prefix the command by the words ‘by elect:’ (Note the colon) and press RETURN.

If necessary replace ‘by elect’ with ‘bysort elect:’

Now type the following commands into your command window (except for help commands which should be typed into the help viewer), pressing return after each. When the command works properly copy it and paste it into your do-file window.

```
summarize turnout if elect==1989
```

```
drop if country=="nir" /*Northern Ireland was sometimes a separate country */
```

Note use of /* comment */ markers. Useful in Do files. A line beginning with * is the same as a line enclosed in /* */

```
table elect, c(mean turnout)
```

Use 'help table' to find out about the various kinds of tables in STATA

HOMEWORK EXERCISE

help crosstab

Use crosstabs to find out what is the proportion of compulsory voting countries in each year. Does the proportion go down over time?

HOMEWORK EXERCISE:

help regress

Do a regression of turnout on compuls timeleft first

→regress turnout compuls timeleft first if elect<2004

FOR THURSDAY, AFTER GOING THROUGH THE EXERCISES...

→gen lowturnout=0

→replace lowturnout=1 if pol==1 | sol==1 | cze==1 | sle==1

A quicker way to do this, especially if there are a lot of replacements to make...

→for var sla pol cze sle: replace lowturnout=1 if X==1

For X in num 1999 2004: gen yX=0

[Here X is a 'stand-in' variable, standing in for the strings '1999' and '2004']

For num 1999 2004: replace yX=1 if elect==X

[Note that X is assumed if no other stand-in variable is named]

[STATA 9 and later has a new 'forevery' command that is much more versatile but not as snappy to use]

Setting memory and matsize [On the Mac this is done in Prefs → Set general prefs; on a Windows machine this is done using 'set memory' and 'set matsize']

Arranging windows for maximum convenience

[STAT TRANSFER for moving data from SPSS]

[ESTTAB for creating tables from regression analysis]

TYPE 'help updates'