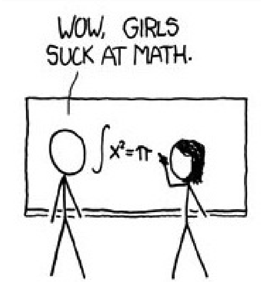
Estimation



Statistical Reasoning
and Quantitative Methods

François Briatte & Ivaylo Petev

Session 5

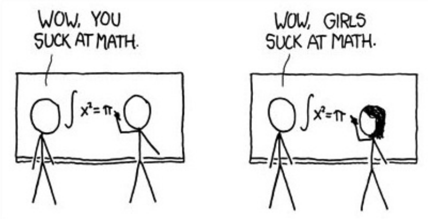
Outline

The properties of the standard normal distribution allow for statistical inference: the estimation, at a certain level of confidence, of the unobserved population parameters, using observed sample parameters.

But first...

Assignment No. 1

Estimation



Assignment No. 1

Univariate statistics

- Introduction
- Datasets
- Distributions
- Estimation

Assignment No. 1

Bivariate statistics

- Significance
- Crosstabulation
- Correlation
- Linear regression

Assignment No. 2

Statistical modelling

- Basics
- Extensions
- Diagnostics
- Conclusion

Final paper

How to proceed

First of all, complete the projects table: `http://goo.gl/brYmB`

This table has to be complete for grading purposes. It has to include your names, class day and time, topic and dataset.

Then:

1. Write your **research design** (using the paper template).
2. Write up your do-file and export your **summary statistics**.
3. **Replicate** your do-file and finalize your paper around 3 to 5 pages.

Finally, email us your work, copying your partner to the email, with the email subject "SRQM: Assignment No. 1, Briatte and Petev" (substitute your own names). The deadline is our next meeting.

Step 1: Research design

Start a text document in about 2 to 4 paragraphs:

- **Topic:** describe your empirical problem in the form of a research question and hypotheses.
- **Data:** describe your dataset with its source, sample size, sampling strategy and variables.

You can revise your hypotheses and selection of variables in future drafts. For now, stick to writing only a few paragraphs, aiming at concision (short and precise descriptions).

The Stata Guide covers all necessary steps in Sections 1–9, with additional formatting instructions in Section 13 and a summary of instructions for this draft in Section 14.

Topic and dataset

- Your **topic** formulates a research question and derives it into testable hypotheses between your dependent and independent variables. Cite previous research if you know any.
- Your **dataset** is a fully referenced and documented sample for which you have read the documentation to understand what each variable measures.

Note that the total **number of observations** on which you will work at later stages might decrease because of missing data, so make sure that you are initially working on the largest possible sample.

Use the `count`, `fre` and `su` commands to inspect missing data, and ignore variables for which there is an excess of missing data.

Variables and description

- Your **variables** have been fully described, as well as renamed and/or recoded if necessary.
- Your **description** of the variables is a summary statistics table with additional observations in your text.

Note that the **number of variables** in your project is determined by your research design, and might evolve if you revise your hypotheses.

For example, you might select something like 6 to 12 independent variables and later focus on only some of them, or even add new ones. You will also be able to revise your do-file accordingly.

Step 2: Summary statistics

Continue with an additional 2 to 4 paragraphs:

- **Summary statistics:** describe all variables in a few words.
- **Normality:** assess how normal your dependent variable is, and describe its potential transformation.

For tables and graphs:

- The **summary statistics table** has to describe *both* your continuous and categorical variables. The table is required to appear in your paper.
- The **histogram** that shows the distribution of the dependent variable can be completed by other plots showing interactions with independent variables.

The tsst command

```
* Short example.
use datasets/nhis2009, clear
* Quick help.
tsst using test.txt
* Quick test.
tsst using test.txt, su(age) fr(sex health) replace
```

Run `draft1.do` for a full example. Look at formatting instructions in the Stata Guide, Section 13.4, which includes an alternative to `tsst` if the command does not work. You can also load it manually:

```
* Run this line before the example above if tsst fails.
run programs/tsst.ado
```

Additional plots

- The **histogram** and **diagnostic plots** that show the distribution and (ab)normality of your dependent variable are required. Also try the `gladder` command to find any potential transformation.
- **Box plots** (`gr box` and `gr hbox`), **dot plots** and **bar plots** (`gr dot` and `gr bar`) with the `over` option can also be used to split your dependent variable over categorical independent variables.
- **Spline plots** with the `spineplot` command are also recommended if you have categorical independent variables to cross-visualize.

Your plots appear only in your do-file, except if you have *specific* observations to make about them. In your paper, include 0 to 2 plots.

Step 3: Replication

Assignment No. 1 consists of:

- **A short paper** named in the format `Briatte_Petev_1.pdf`.
- **A short do-file** named in the format `Briatte_Petev_1.do`.

The do-file will include all that is needed to replicate your results: subsetting, variable renaming and recoding, summary statistics, normality tests, and tabular and graphical visualizations.

Replicate your do-file before sending it, to make sure that it executes properly and produces the results displayed in your analysis.

Further help

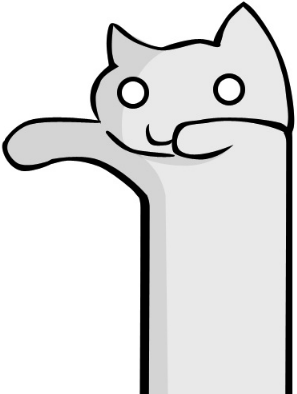
- Course-specific help:
  - Stata Guide
  - Session do-files
  - Course slides
- General help:
  - Handbook chapters
  - Stata documentation (`help` *`command`*)
  - Online tutorials

Everything is systematically archived on the course website:

`http://f.briatte.org/teaching/quanti/`

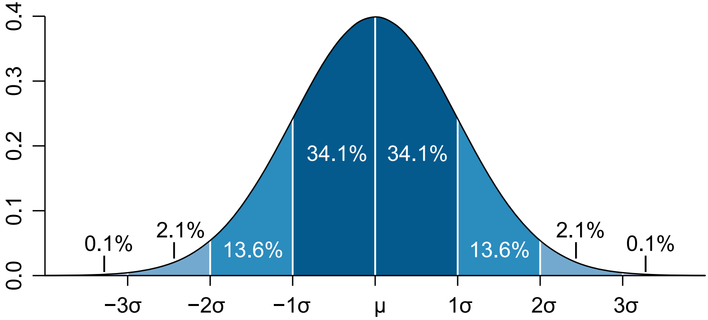
Happy coding!

And now, estimation.



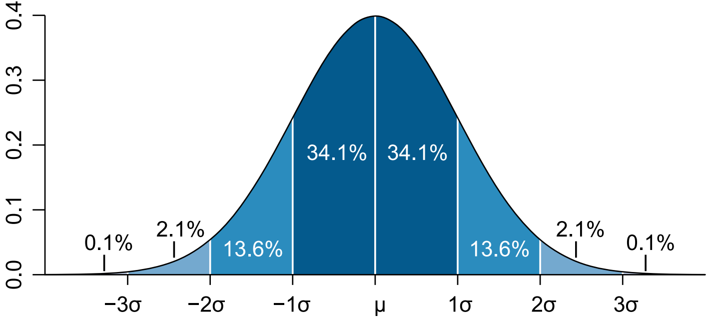
Standard normal distribution $\mathcal{N}(0, 1)$

- $\mu \pm 1\sigma$ contains approximately **68%** of all values.
- $\mu \pm 2\sigma$ contains approximately **95%** of all values.
- $\mu \pm 3\sigma$ contains approximately **99%** of all values.



Probability density function

- $Pr(\mu - 1\sigma < \mu < \mu + 1\sigma) \approx .68$
- $Pr(\mu - 2\sigma < \mu < \mu + 2\sigma) \approx .95$
- $Pr(\mu - 3\sigma < \mu < \mu + 3\sigma) \approx .99$

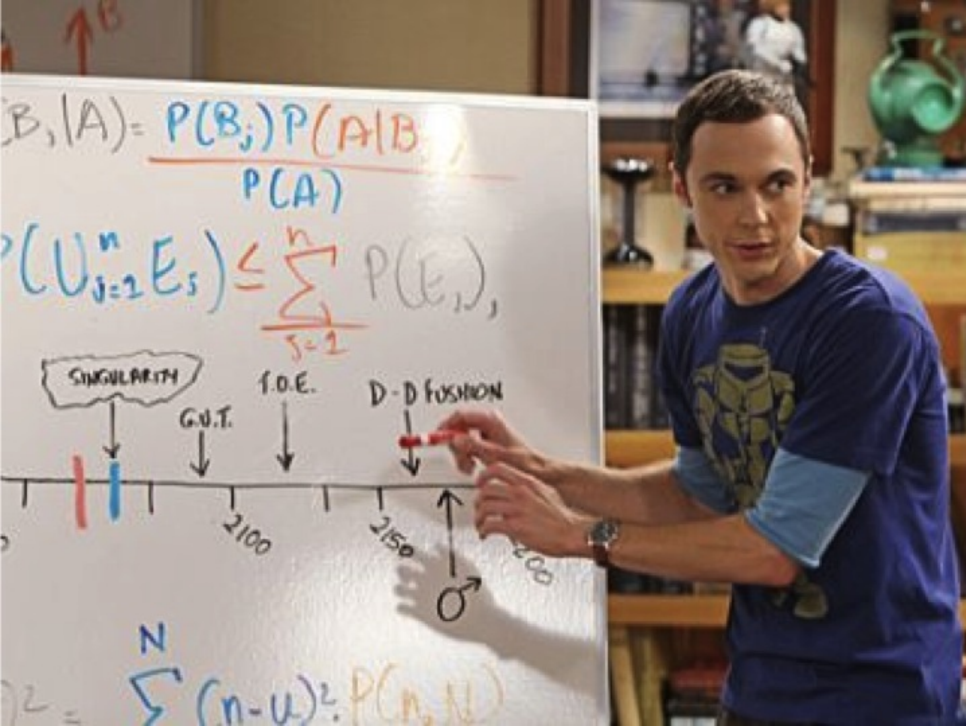


Estimation

The normal distribution is used for statistical inference, i.e. for estimating the population parameters from the sample parameters.

Parameter	Notation	
	Sample	Population
Mean	$\bar{X}$	μ
Standard deviation	s	σ

This operation generalizes the values held by the sample to the population under consideration, under certain levels of confidence.


P(A)
SINGULARITY
T.O.E.
D-D FUSION
G.U.T.
2100
2150
N

Central Limit Theorem

Formally, μ is the population mean, which is unobserved, and $\bar{X}$ is the sample mean, which is observable by analysing the data.

The **Central Limit Theorem** (CLT) states that, if repeated samples are drawn from the population, their respective means $\bar{X}_1, \bar{X}_2, ..., \bar{X}_n$ will be normally distributed around the population mean μ:

$$\text{CLT} : \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\bar{X}_i - \mu\right) \xrightarrow{d} \mathcal{N}(0,\ \sigma^2)$$

The CLT holds regardless of the distribution of the variable X under examination, which is, like, totally awesome.

Law of Large Numbers

Formally, σ is the population standard deviation, which is unobserved, and s is the sample standard deviation, which is observable by analysing the data.

The **Law of Large Numbers** states that, the larger the sample, the closer its mean will be to the population mean. The standard error of that estimate derives from the standard deviation of the variable.

The standard error of the mean (SEM), which estimates the standard deviation of the population from the standard deviation in the sample, will hence (slowly) decrease with sample size:

$$\text{SEM} = SE_{\bar{X}} = \frac{s}{\sqrt{N}} \text{ (sample)}$$

Confidence intervals

Given these properties, the population mean μ can be estimated from a sample X of N (statistically independent) observations.

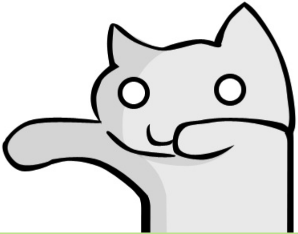
When we observe a normal distribution, $\mu \pm 2\sigma$ contains approximately 95% of all values. The exact number used for estimation at 95% confidence, called the z-score, is $z = 1.96$.

When the sample values of X are normally distributed, $\bar{X} \pm 1.96 \cdot SE_{\bar{X}}$ contains 95% of the possible values of μ.

These bounds define a 95% confidence interval:

$$\bar{X} - 1.96 \cdot SE_{\bar{X}} < \mu < \bar{X} + 1.96 \cdot SE_{\bar{X}}$$

- In 2.5% of cases, $\mu < \bar{X} - 1.96$.
- In 2.5% of cases, $\mu > \bar{X} + 1.96$.



Stata implementation

Summary statistics for current worldwide fertility rates:

```
su births
```

Variable	Obs	Mean	Std. Dev.	Min	Max
births	186	3.138294	1.649826	1.1	7.446

Standard error and 95% confidence interval:

```
ci births
```

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
births	186	3.138294	.1209711	2.899634	3.376954

- Use the `level(99)` option for a larger 99% confidence interval.
- Use the `binomial` option if the variable is dichotomous (binary).

Stata implementation: Q&A

- **Q: Why is the 99% confidence interval larger?**

  A: The 99% CI uses $z = 2.58$ to include 99% of all possible values of μ in the interval. Its calculation with $\bar{X} \pm 2.58 \cdot SE_{\bar{X}}$ therefore includes more values on both sides of $\bar{X}$.

  This is called the precision-accuracy trade-off: higher confidence for μ is obtained at the expense of a more precise value for μ. Maximising sample size will attenuate the trade-off.

- **Q: Why do we apply a 'binomial' option to binary variables?**

  A: Binary outcomes of "yes/no" variables form a distribution with no intermediate values that cannot be normal. Instead, the discrete probability of independent 0/1 outcomes (Bernoulli trials) is given by the binomial distribution.

  We will focus on estimation based on the normal distribution.

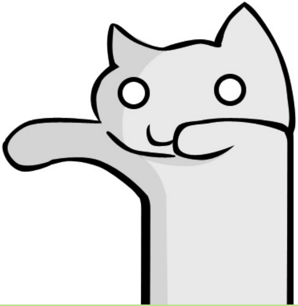
Back to estimation

Using the normal distribution as a probability density function and standardised *z*-scores to select a level of confidence:

	Notation	
Parameter	Sample	Population
Observations	N	unobserved
Mean	$\bar{X}$	μ
Standard deviation	s	σ
Standard error	SE	

Estimating μ at 95% or 99% confidence:

$$Pr(\mu \in \bar{X} \pm z = 1.96 \cdot SE_{\bar{X}}) = .95$$
$$Pr(\mu \in \bar{X} \pm z = 2.58 \cdot SE_{\bar{X}}) = .99$$



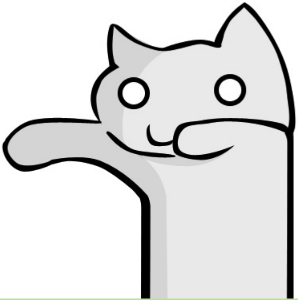
Back to estimation

For a variable X, estimating μ from $\bar{X}$ relies on three prerequisites that you will apply to all your variables:

- **Assess the distribution**'s normality i.e. whether $X \sim \mathcal{N}(\mu, \sigma)$.
- **Select the standardised** z**-score** to fix a level of confidence.
- **Minimise the sampling error** by maximising the sample size (N).

This technique guides all point estimation that we perform in this course. It is called Maximum Likelihood Estimation (MLE).

Following the assumptions of MLE pleases **Estimation Cat**, the first of two cats in The Prophecy.



The Prophecy

The Prophecy states that the powers of Estimation Cat (white) are limited by those of **Significance Cat** (black).

- **Estimation** provides a parameter and its standard error. Relationships between variables can be modelled as parameters.
- **Statistical significance** provides its probability level, i.e. the probability that the estimated parameter is different from zero.

Parametric statistics work by confronting the awesome powers of Estimation Cat and Significance Cat into a model.

This course is an introduction to statistical modelling using frequentist statistics, a.k.a The Prophecy.