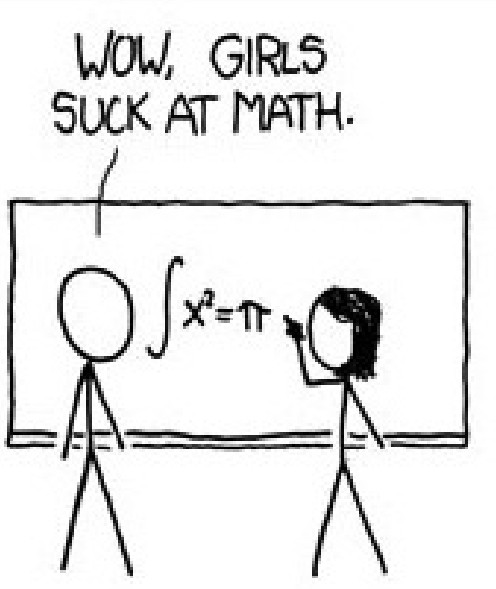
Estimation



Statistical Reasoning and Quantitative Methods

François Briatte & Ivaylo Petev

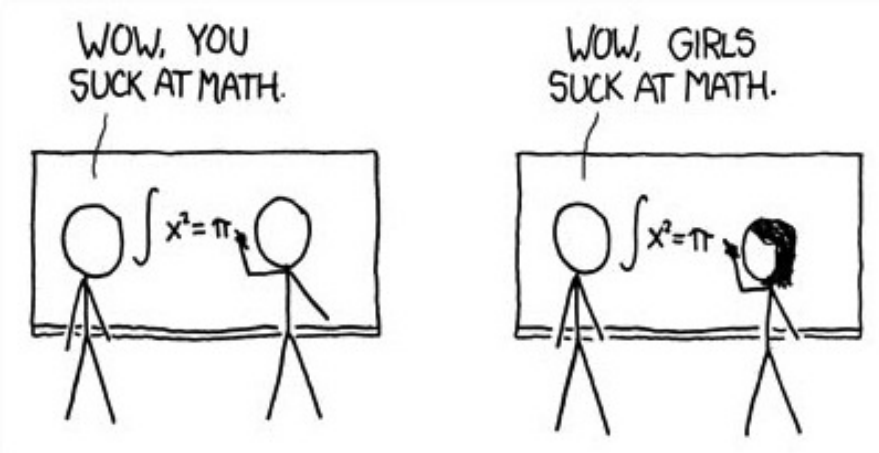
Session 5

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.

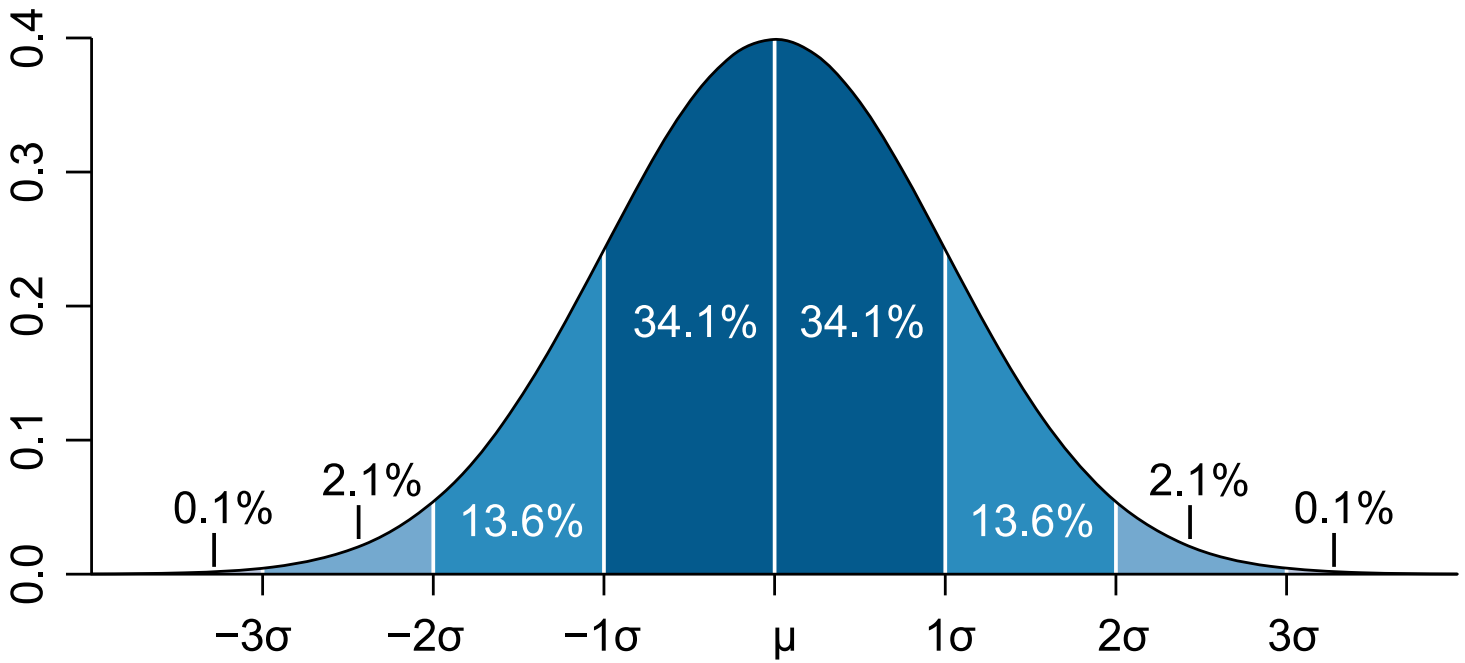
Estimation

Assignment No. 1



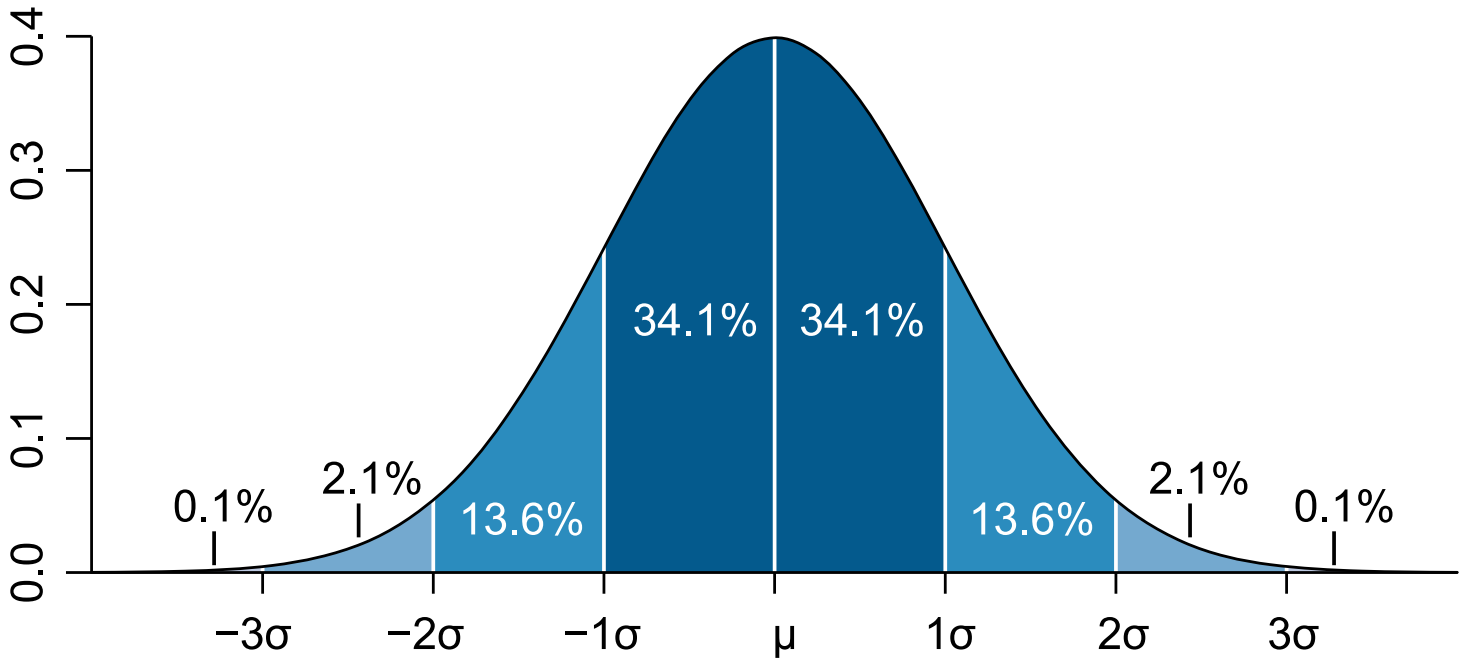
Standard normal distribution $\mathcal{N}(0, 1)$

- $\mu \pm 1\sigma$ 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

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

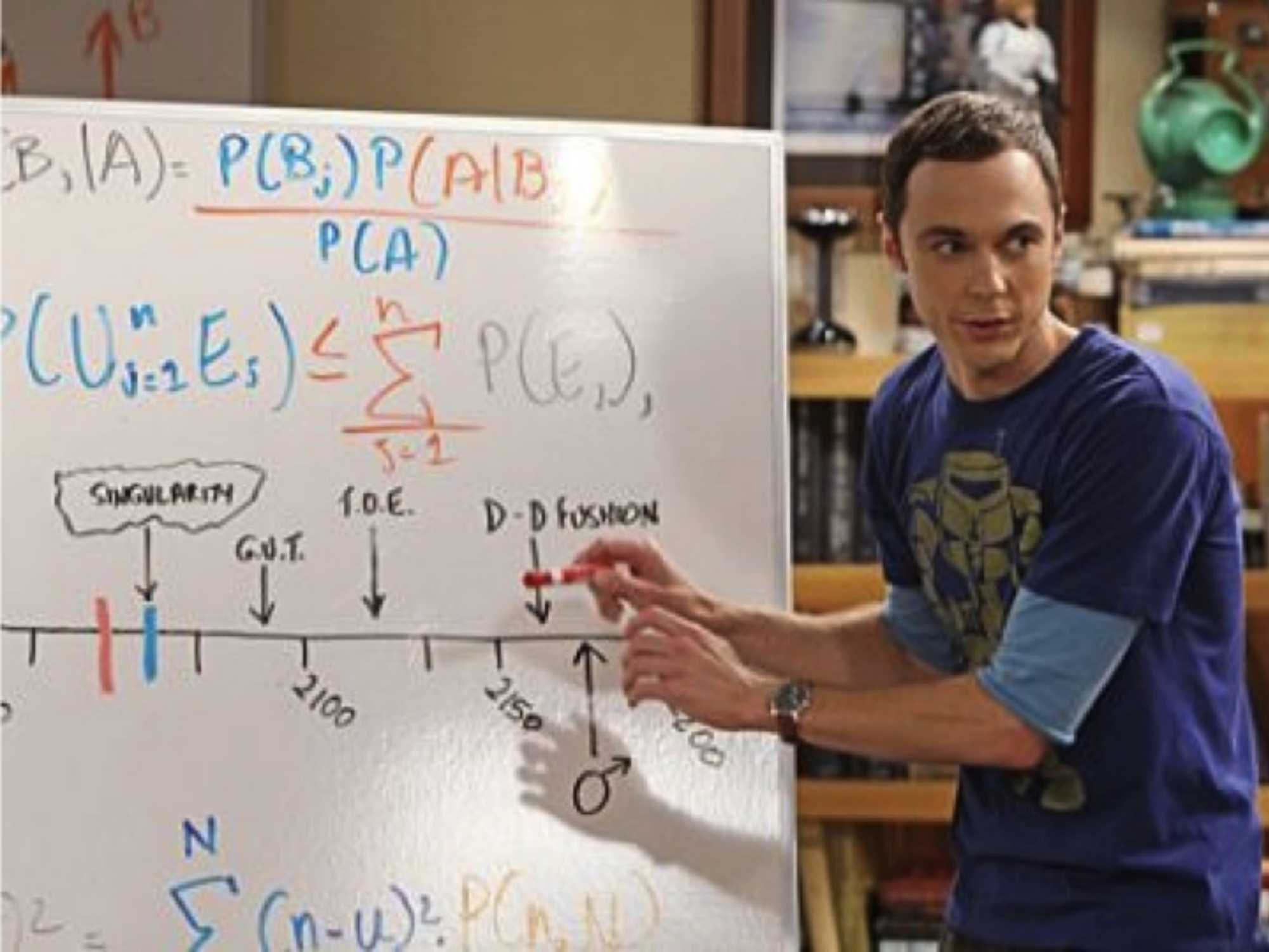


Estimation

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

	Notation	
Parameter	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.


$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{P(A)}$
$P(\bigcup_{j=2}^{n} E_j) \leq \sum_{j=2}^{n} P(E_j)$,
SINGULARITY
G.U.T.
T.O.E.
D-D FUSION
2100
2150
$\sum_{n=1}^{N} (n-u)^2 \cdot P(n,N)$

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.

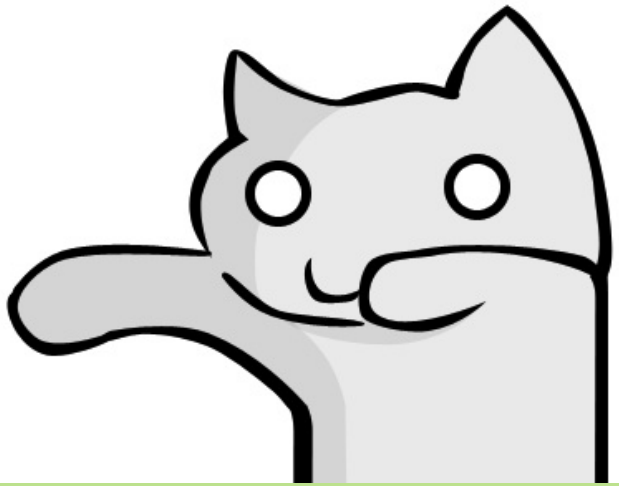
When we observe a normal distribution, $\mu \pm 2\sigma$ contains approximately 95% of all values. The exact number used for estimation at 95% confidence, called the z-score, is $z = 1.96$.

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.

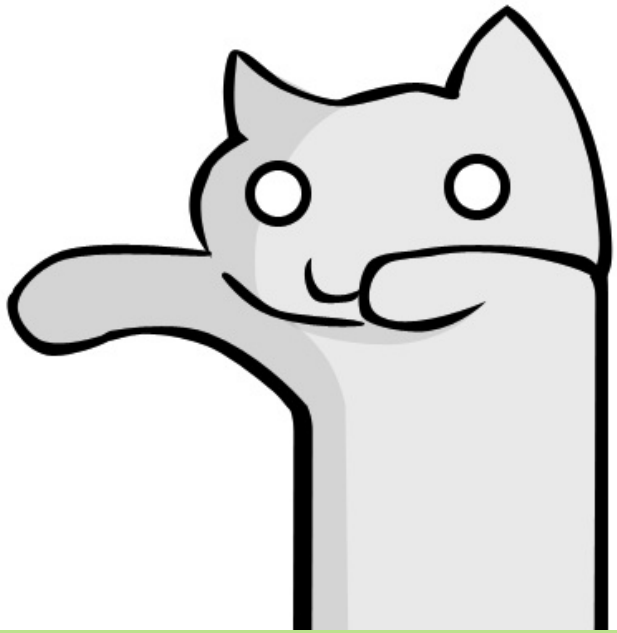
Back to estimation

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:

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



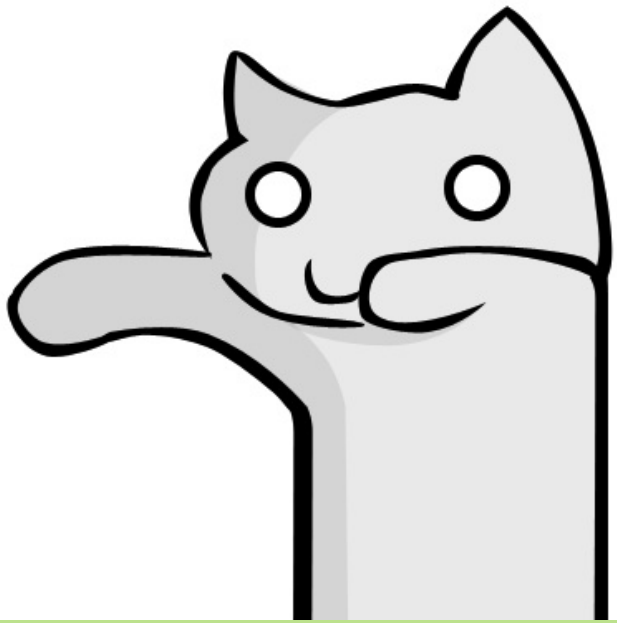
Back to estimation

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.



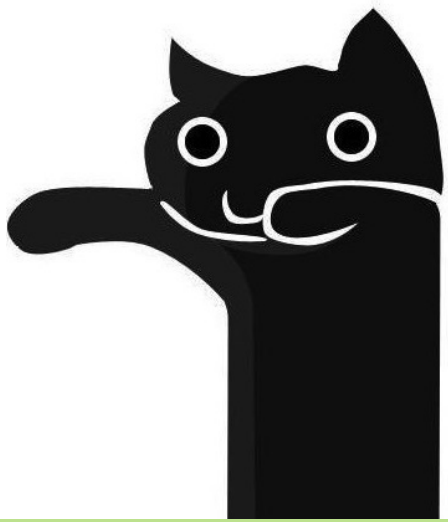
The Prophecy

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.



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

Things you should already have done:

- **Read the course planning**: `http://goo.gl/BJHkQ`
- **Discuss your project**: `http://goo.gl/brYmB`
- **Look at a template**: `http://goo.gl/7u8oa`

(Copied from Session 3.)

The course planning mentions the deadline, the list of projects gives you an idea of what is going on, and the template shows you how to structure your paper. Now, for the draft itself:

1. Write your **research design** (using the template).
2. Export your **summary statistics** (from the do-file).
3. **Replicate** and send us your work (copy your partner).

Step 1: Research design

Start a text document with two 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.

In your final paper, you will develop your topic into an "Introduction" with revised hypotheses, and discuss your data and variables in the "Methods" sections. For now, stick to writing paragraphs.

Refer to the Stata Guide, Sections 9 and 14, for details on this step and the next one, summary statistics. For help on Step 3 (replication and sending files), refer to Sections 1 and 13.

Step 2: Summary statistics

Continue with two more paragraphs:

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

Add tables and graphs:

- The **summary statistics table** to describe all continuous variables is produced using `tabstat` and `tabstatout`.
- The **frequency table** of categorical variables is produced using the `tabout` command, and should be fit into summary statistics.
- The **graphs** describe the distribution of the dependent variable in the sample and across categorical independent variables.

Step 3: Replication

Assignment No. 1 consist 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 the do-file before sending it, to make sure that it runs correctly on your data.

Send both files under the subject "**SRQM: Assignment No. 1, [Last Names]**" to us and to your partner, and you are done!

Please refer to the Stata Guide, Section 13, for details.

Course do-files replication

As compulsory homework, replicate each do-file used in the course sessions, which are archived online:

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

- Download each do-file to your `Replication` folder, and make sure that all course datasets are stored in the `Datasets` folder.
- Open Stata, set your working directory to the **SRQM** folder and adjust other settings as shown in `week1.do`.
- To replicate, type e.g. `doedit "Replication/week2.do"` and run it while reading the comments.

(Copied from Session 3.)

Course do-files replication

The reason why replication is compulsory homework has to do with your learning experience throughout the course:

- **It helps with understanding the course session again** (or with catching up if you skipped class). You can also open the do-file in Stata and read it while replicating its commands sequentially.
- **It saves plain text log files in Teaching Pack ▷ Replication** that will show all comments from the do-file, along with its commands and their results.
- **It familiarises you with using Stata on your personal system** and contains several bits of code like commands, options and structure that can be adapted to fit your own research project.

Also, this is also how real-world quantitative researchers work: by reproducing each other's work, using replication material.

Personal do-files

We will assess your research project by replicating your do-file and evaluating your analysis:

- **Your do-file** needs to use the appropriate commands at the right place, with some structure and comments to guide us through your work. Try replicating your own do-file before submitting it: the code must execute properly and the results must correspond to those reported in your work.
- **Your text** contains the interpretation of (some of) the do-file results: without a proper do-file, you will struggle to produce any accurate analysis. Valid statistical reasoning needs *much* more than a do-file, but correct, comprehensible code is a prerequisite to writing up a coherent quantitative paper.

Real-world replication example

As optional homework, replicate economist Dani Rodrik's paper "**The Real Exchange Rate and Economic Growth**" (2008):

`http://www.hks.harvard.edu/fs/drodrik/research.html`

- Download the paper (PDF) and replication material archive (RAR) from Dani Rodrik's online research page.
- Open Stata and set your working directory (`cd`) to the folder created by decompressing the archive.
- Rename the do-file as `master.do` and run the following command in Stata: `do master.do, nostop`

Now watch the figures pop up and regression models run in the **Results** window. A few uninstalled commands and some of Dani Rodrik's personal settings will be skipped by the `nostop` option.

Real-world replication example

On my own system, I added a few tweaks that you can adapt to your own system if you want to replicate the analysis in full:

- After downloading all files to my Desktop, I renamed the folder to `rodrik2008` and the do-file to `master.do`.
- In the do-file, `graph export` is set to use WMF; on Mac OS X, find and replace "`.wmf`" by "`.pdf`" to export all figures.
- The `outreg2` and `xtabond2` packages need to be installed with `ssc install` to ensure that all commands will run.

The `rodrik2008` folder will contain graphs, exported tables and a log file named `arellano_bond.txt` after the do-file has run.

Further help

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

Handbook chapters and course emails are available from the ENTG. Everything else is systematically archived on the course website:

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

Happy coding!