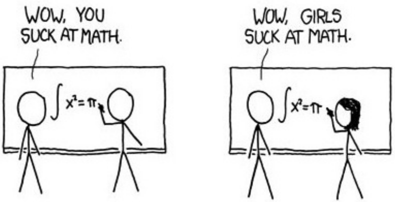
ESTIMATION / DRAFT 1

1 Draft No. 1

2 Estimation

3 Practice



Draft No. 1

Univariate statistics

- Introduction
- Datasets
- Distributions
- Estimation

Bivariate statistics

- Significance
- Comparisons
- Correlation
- Regression

Statistical modelling

- Basics
- Extensions
- Diagnostics
- Conclusion

First draft

Revised draft

Final paper

Instructions

Catch up

- **Readings:** handbook chapters, Stata Guide
- **Replication:** all do-files so far
- **Projects:** full registration required!

Write up

- **Copy** the paper template to your Google Drive
- **Share** the (renamed) template within your group
- **Upload** the draft PDF and do-file to the ENTG

Common mistakes

Project files

- **Filenames:** use your group shortname
- **Formats:** use PDF (paper) and .do (code)
- **ENTG:** use groups 57344, 57345 and 57346

Paper content

- **Paragraphs:** respect limits, write analytically
- **Table formats:** single-paged, rounded figures
- **Sources:** format your bibliography, cite the data

Survival techniques

Code

- **Copy** code chunks from the course do-files
- **Run** the code entirely to check for errors
- **Log** your results to analyze them later

Paper

- **Write** as the authors of the example papers
- **Hypothesize** from structural determinants to covariates
- **Select** only the statistics that you can analyze

Extra tips

Research tips

- **Maximize sample size:** keep the data representative
- **For a continuous DV,** normality matters for linear regression
- **For a categorical DV,** recode to binary for logistic regression

Graph tips

- **Keep graphs open** with `name()`; leave them out of the paper
- **Plot over small multiples** with `over()` and `by()`
- **Use the IOTT** to decide whether to keep or ditch a graph

Grading scheme

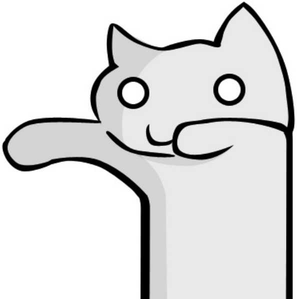
Protocol

1. **Replicate findings:** run code, open results log
2. **Review design:** check sources, hypotheses
3. **Suggest improvements:** variables, plots, references, …

Best papers

- **Questions on front page**, with line numbers for code issues
- **No spelling mistakes**, with no formatting issues throughout
- **Selected, rounded statistics**, all with a specific interpretation

And now, estimation.



Estimation

The issue

- The **sample parameter** is the sample mean $\bar{X}$
- The **population parameter** is the population mean μ
- How to **generalize** from sample to population?

The solution

- **Central Limit Theorem** (CLT)
- **Law of Large Numbers:** (LLN)
- **Confidence intervals:** (CIs)

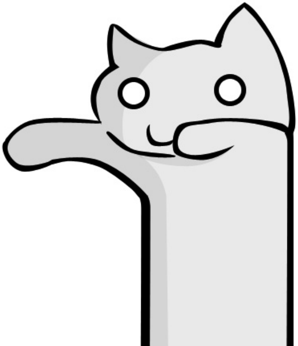
Central Limit Theorem

Definition

The independent and identically distributed means $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_n$ of repeated random samples are normally distributed around μ.

Formula

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



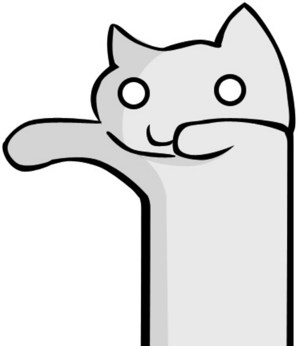
Law of Large Numbers

Definition

The sample standard deviation SD_x converges towards the population standard deviation σ at a speed of $\sqrt{N}$.

Formula

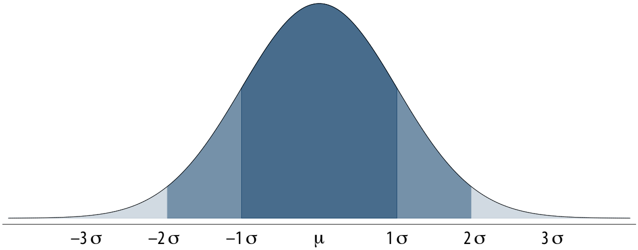
$$\mathrm{SEM} = SE_{\bar{X}} = \frac{SD}{\sqrt{N}}$$



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

Properties

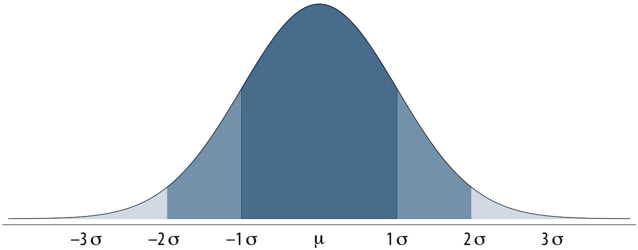
- $\mu \pm 1\sigma$ contains approximately **68%** of all values.
- $\mu \pm 2\sigma$ contains approximately **95%** of all values.
- $\mu \pm 3\sigma$ contains approximately **99%** of all values.



Probability density function

Properties

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



Confidence intervals

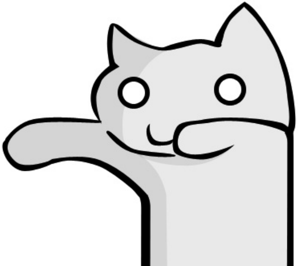
Given these properties, the population mean μ can be estimated from a sample X of N (statistically independent) observations.

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.

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



Practice: NHIS dataset

$$\text{Body Mass Index} = \frac{\text{mass (kg)}}{(\text{height(m)})^2} = \frac{\text{mass (lb)} \times 703}{(\text{height(in)})^2}$$

- For **normal weight** adults, $18.5 < \text{BMI} < 25$.
- For **overweight** adults, $25 \leq \text{BMI} < 30$.
- For **obese** adults, $\text{BMI} \geq 30$.

Data:

- National Health Interview Survey (NHIS)
- Sample: U.S. adult population, 2009

Practice session

Class

```
* Get the do-file for this week.
srqm fetch week4.do
* Open to read and replicate.
doedit code/week4
```

Coursework

- Finish the do-file and read all comments at home.
- Follow instructions on top of the code.
- Prepare questions in your group's draft do-file.

Exercise

Ex 4.1. Quality of Government 2011

1. What countries have *much* more females in government?
2. How is the female-to-male income ratio distributed?
3. Same question with confidence variables (`d wvs_e069*`).
4. Plot the Gini coefficient over quartiles of GDP per capita.

Tips

- Label outliers: `gr hbox x, mark(1, mlab(ccodealp))`
- Get quartiles: `xtile qx = x, nq(4)`