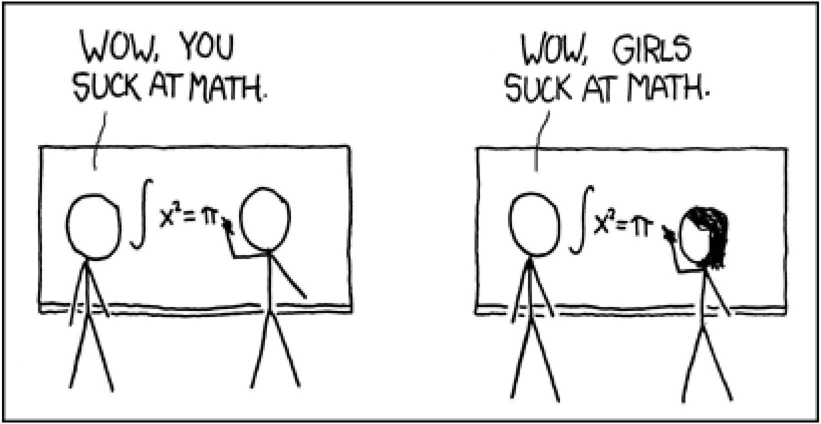
Statistical Reasoning

Week 5

Sciences Po - Louis de Charsonville

Spring 2018


WOW, YOU
SUCK AT MATH.
WOW, GIRLS
SUCK AT MATH.

Outine

Research Paper

Standard errors & Central Limit Theorem

- Definition
- Central Limit Theorem
- z-scores & t-values

Research Paper

Research Paper

Timeline

1^{st} draft	**Today. Deadline : 23h59**
2^{nd} draft	**10 April**
Final draft	**24 April**

Submission's Rules

- A **word** document (following template on the Google Drive).
- A **do-file** showing *all* commands in Stata with comments in green.

Standard errors & Central Limit Theorem

Standard errors

- People usually gloss over this abstract concept.
- This is a **huge mistake**.
- It is the denominator of numerous formulas to compute inferential statistics.

Standard Error

A **standard error** is the standard deviation of the *sampling distribution*.

- Read example p.49 of Urdan (2010) *Statistics in Plain English*.

Example - from Urdan (2010) 1/2

We want to find the average shoe size of all adult American women.

1. Select a *sample* of 100 women at *random*.
   - Our sample may or may not look like the typical American women. *But* any difference is due to chance.
2. Compute the average shoe size for this sample.
3. Throw the first sample and draw again a sample of 100 women.
4. The second sample may have an average shoe size that is quite different from our first sample.
5. Do again step 1 & 2 a thousand times.
6. ⇒ The collection of samples' average is the **sampling distribution**
7. Compute the standard deviation of the sampling distribution. This is the **standard error**.

Mean and standard deviation of the sampling distribution

- Recall : *standard deviation* is the *average difference or deviation from the mean*.
- The mean of the *sampling distribution* is called the **expected value of the mean**

Standard error represents :

- Average difference between the expected value (e.g. population mean) and an individual sample mean.
- How confident we should be that a sample mean represents the actual population mean.
- A measure of how much error we can expect when we approximate the mean of the population by a sample's mean.

How to calculate the Standard Error 1/2

- Background : shoe sizes from a sample of 100 American womens
- Sample average : 6
- Best guess : this is the population's average.

How much *error* can I expect from this *guess* ?

How to calculate the Standard Error 1/2

- Background : shoe sizes from a sample of 100 American womens
- Sample average : 6
- Best guess : this is the population's average.

How much *error* can I expect from this *guess* ?

Intuitions

- How large is my sample ?
- How much variation in my sample (e.g. standard deviation) ?

How to calculate the Standard Error 2/2

Standard Error of the Mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{1}$$

or

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{2}$$

where $\sigma =$ the standard deviation for the population

$s =$ the sample estimation of the standard deviation

$n =$ the size of the sample

Central Limit Theorem 1/2

Central Limit Theorem - in plain English

For a large sample size (e.g., $n = 30$), the sampling distribution of the mean will be normally distributed, even if the distribution of scores in your sample is not.

Central Limit Theorem - mathematically

Let $\{X_1, \ldots, X_n\}$ be a sequence of n independent and indentically distributed random variables drawn from distributions of expected values μ and finite variances σ^2.
Let S_n be the sample average :

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{3}$$

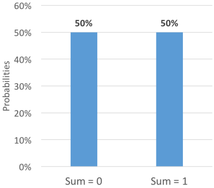
Then as n tends to the infinite, S_n converges towards a normal distribution :

$$\frac{S_n - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \mathcal{N}(0,1) \tag{4}$$

Example 1 - Coin flip Let's flip a coin. We have a head with probability $1/2$ and a tail with probability $1/2$.

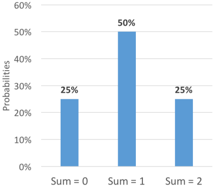
- For 1 coin flip

Rank	Outcome	Sum of outcomes	Probability
1^{st}	0	0	50%
	1	1	50%



- For 2 coin flips

Rank	Outcome	Sum of outcomes	Probability
1^{st}	0	$0+0=0$	25%
	1	$0+1=1$	50%
2^{nd}	0	$1+0=1$	
	1	$1+1=2$	25%



Intuition 2/2

Example

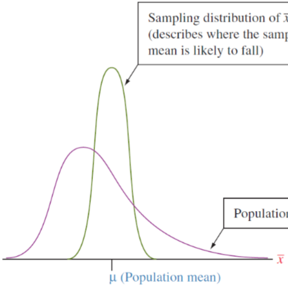
http://louisdecharson.github.io/temp/central_limit.html

Main features of Central Limit Theorem

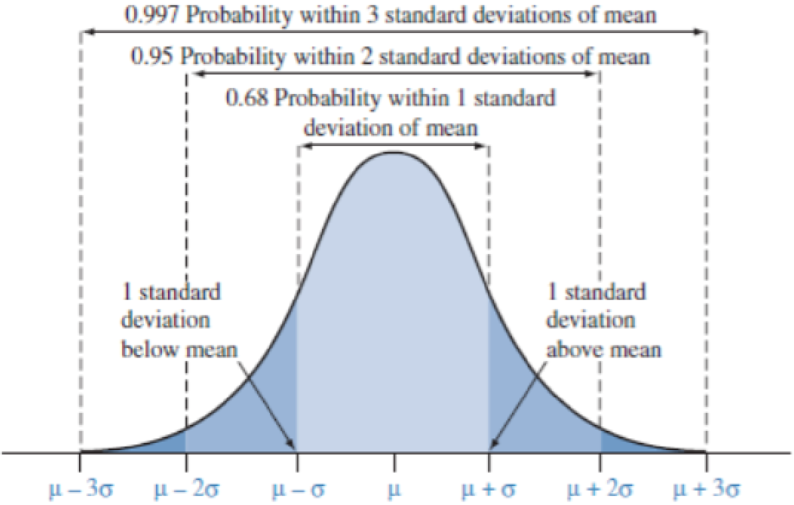
- Whatever the distribution of thesample or of thepopulation, if the sample size is large enough (n greater than 30), the theoretical sampling distribution will have a normal distribution ;
- The mean of the sampling distribution is the population mean (i.e. the unknown parameter we want to find out) ;
- The standard deviation of the sampling distribution indicates the range of possible error between a given sample mean and the population mean, or what is called the standard error.

Implications

- We do not know where our sample mean is located in the sampling distribution ...
- ... But because the sampling distribution is normally distributed ...
- We know that it is very improbable that our sample mean is not in +/- 3 standard errors from the true parameter (= the population mean = what we want to know).



Central Limit



Standard Error

Recall : the standard error is the estimated standard deviation of the sampling distribution :

$$se = \frac{sd}{\sqrt{n}}$$

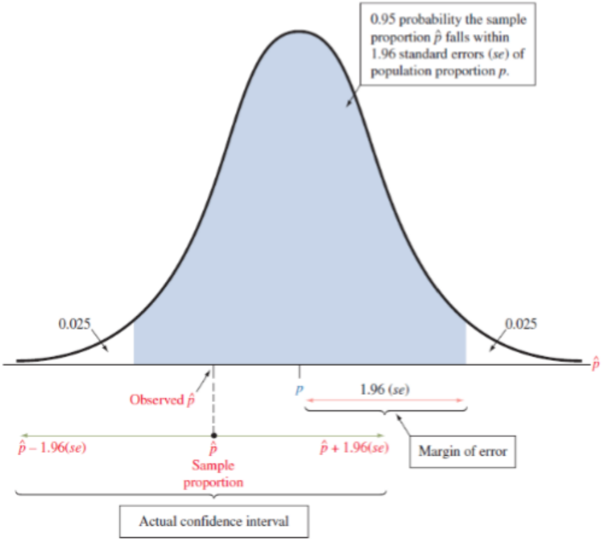
with se = standard error and sd = standard deviation of our sample.

- the **larger** the sample, the **lower** the standard error ;
- the **larger** the standard deviation in our sample, the **larger** the standard error.

Confidence Interval 1/3

- What we are interested in : the parameter of the population
- What we can do : *estimate* it only (it will remain *unknown*)
- This estimation consists of :
  - a **point estimate**
  - a **confidence interval** : an interval within the parameter value is believed to fall with a certain degree of confidence / probability. Most of the time, the confidence level chosen is 95%

- To calculate the confidence interval based on a confidence level of 95%, we use a feature of the normal distribution :
- We know that when data are normally distributed, 95,44% of the observations fall within 2 standard deviations of the mean.
- A exact probability of 95% of observations correspond to 1,96 standard deviations of the mean.
- Here, as we are working on a sample, we use the estimated standard deviation of the sampling distribution of the mean, (which is a normal distribution), i.e. the standard error ; As a consequence, a 95% confidence interval for the parameter to estimate (mean of the population) corresponds to the sample mean +/- 1,96 x Standard error ;
- The distance $1.96\,s.e.$ is called the **margin of error**.



0.95 probability the sample proportion $\hat{p}$ falls within 1.96 standard errors (se) of population proportion p.
0.025
0.025
$\hat{p}$
Observed $\hat{p}$
p
1.96 (se)
$\hat{p} - 1.96(se)$
$\hat{p}$
Sample proportion
$\hat{p} + 1.96(se)$
Margin of error
Actual confidence interval

- A z score is a number that indicates how far above or below the mean a given score in the distribution is in standard deviation units.

$$
\begin{aligned}
z &= \frac{\text{raw score - mean}}{\text{standard deviation}} \\
&= \frac{X - \mu}{\sigma} \\
&= \frac{X - \bar{X}}{s}
\end{aligned}
$$

$$
\begin{aligned}
\text{where } X &= \text{raw score} \\
\mu &= \text{population mean} \\
\sigma &= \text{population standard deviation} \\
\bar{X} &= \text{sample mean} \\
s &= \text{sample standard deviation}
\end{aligned}
$$

- z scores tell researchers instantly how large or small an individual score is relative to other scores in the distribution.
- Example : if a students got a z-score of 2 on an exam, it means that the student score 2 standard deviations above the mean on that exam.
- a z-score of 1.96 means that if the distribution is normally distributed the student is in the top 2.5%.

The isssue of sample size

- Remember :

$$s.e. = \frac{sd}{\sqrt{n}}$$

- When n is small, the standard error could be sizeable.
- Moreover, when n is small, the sampling distribution can differ from the normal distribution
- In this case, instead of using the normal distribution, we use the **t Distribution**, which differs according to the *number of degrees of freedom* ($df = n - 1$).
- When $n \rightarrow \infty$, the t Distribution converges toward the normal distribution.

Student's t

- The t distribution has slightly different features compared to the Normal Distribution for small n, which implies a different calculation of confidence intervals.
- In practice, it means that when n is small, the number 1,96 is not appropriate anymore to find 95% of units of the distribution. This number changes according to n (or to the number of degrees of freedom).
- To know which number should be used to calculate a 95% confidence interval, we use the Student's t table.
- In practice, software almost always use Student's t to calculate confidence intervals.