

# Analyse sociologique de données d'enquête par questionnaire

Olivier Martin

Cours de l'UE « Méthodes d'investigation sociologique »

L3 de sciences sociales – Université Paris Descartes – 2011-2012

Séance 8 :

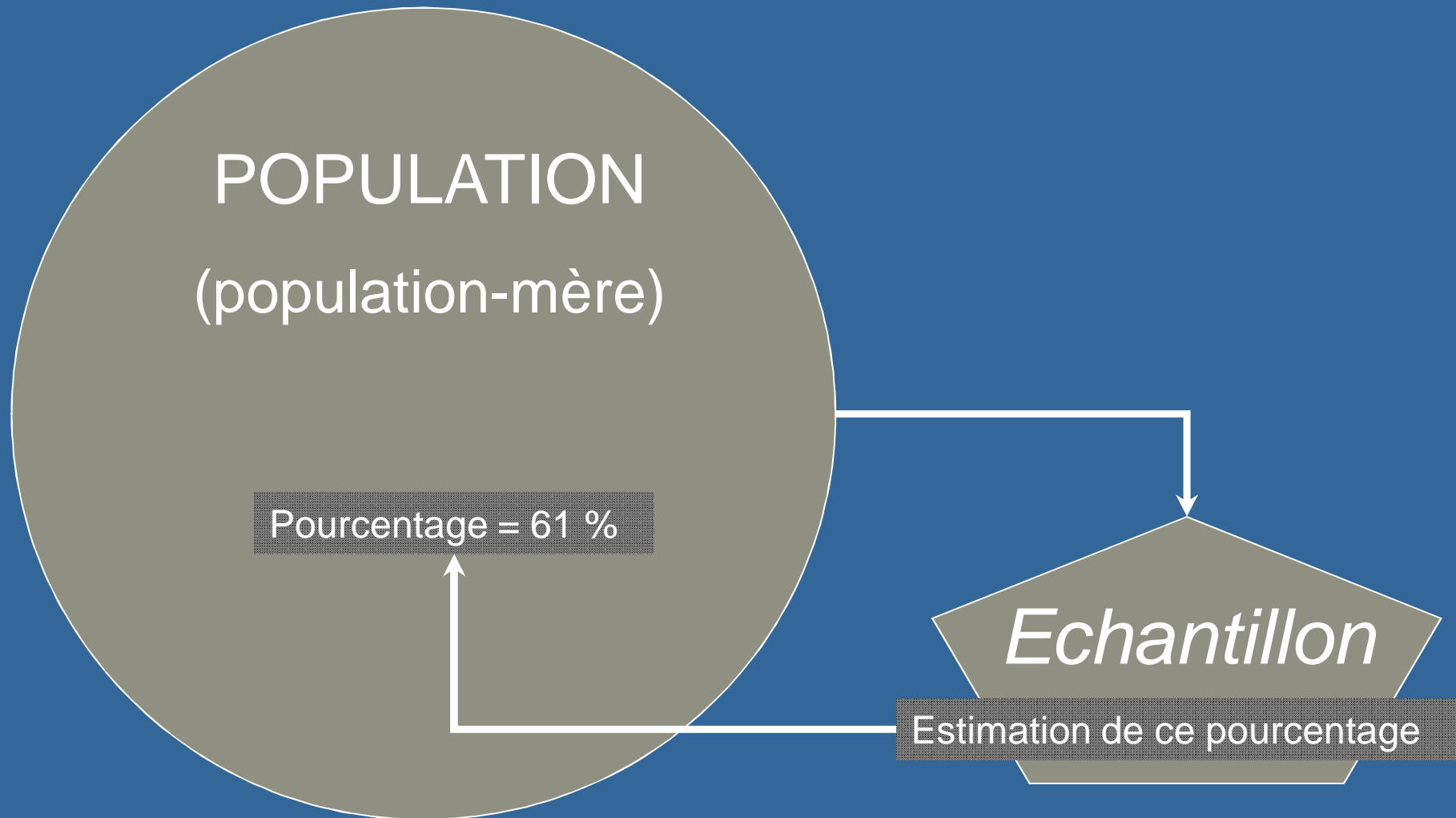
Comment mesurer  
la « fiabilité »  
des pourcentages ?

SUITE

## Démarches « empiriques »

Imaginons la situation suivante :

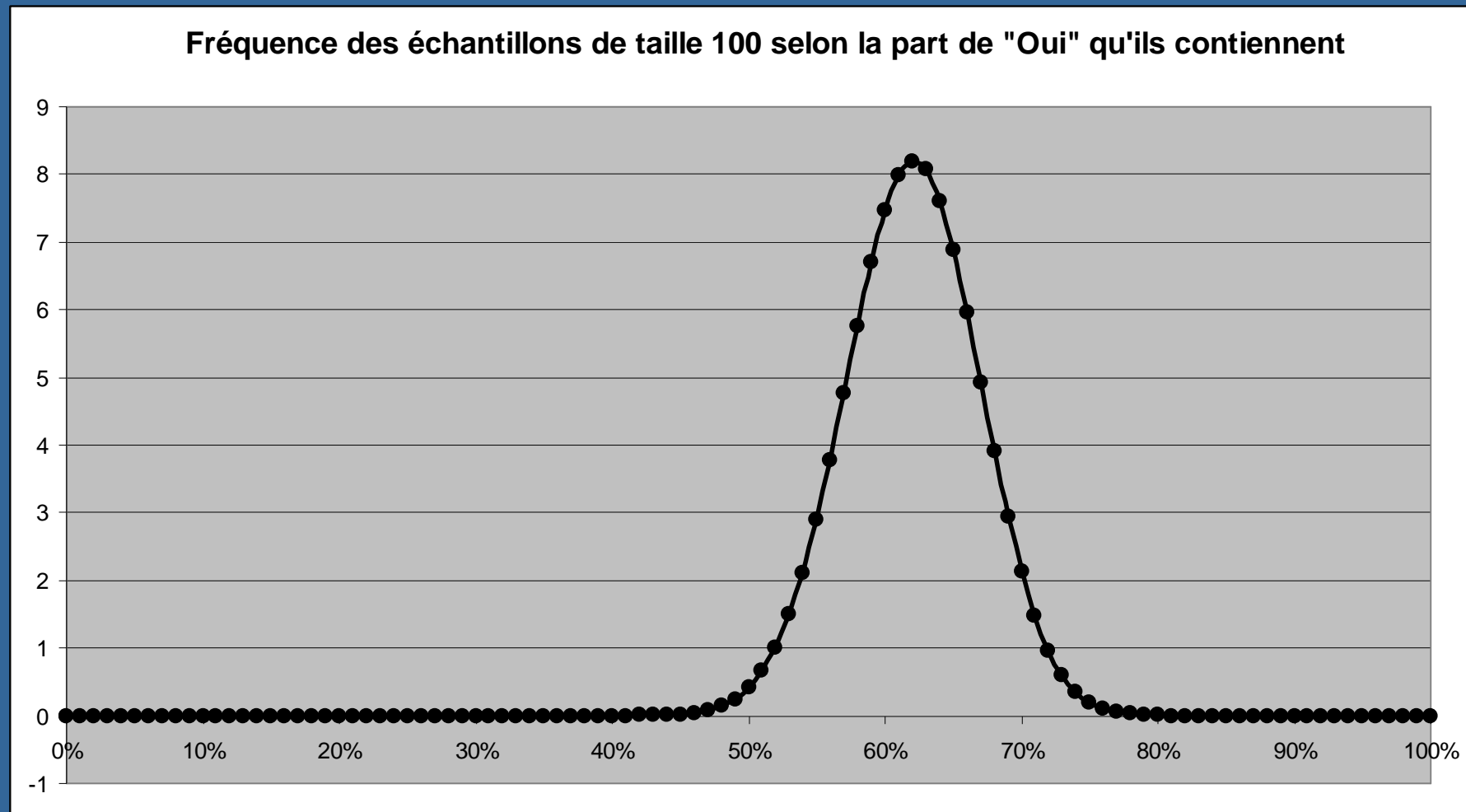
- Population de très grande taille
- On suppose que 61 % de cette population possède la propriété Y (avoir fait telle chose, avoir telle caractéristique...) : 61 % de « Oui » ; 39 % de « Non »
- Quelle estimation de ce pourcentage de « Oui » va nous fournir notre échantillon ?



## Troisième cas : un échantillon de taille 100

- On peut simuler...
- On peut également calculer (calcul de probabilité – dénombrement)
  - quelle est la probabilité d'obtenir un échantillon contenant aucun « Oui » ?
  - quelle est la probabilité d'obtenir un échantillon contenant un seul « Oui » ?
  - ...
    - quelle est la probabilité d'obtenir un échantillon contenant 60 « Oui » ?
  - quelle est la probabilité d'obtenir un échantillon contenant 61 « Oui » ?
  - quelle est la probabilité d'obtenir un échantillon contenant 62 « Oui » ?
  - ...
    - quelle est la probabilité d'obtenir un échantillon contenant 99 « Oui » ?
    - quelle est la probabilité d'obtenir un échantillon contenant 100 « Oui » ?

## Troisième cas : un échantillon de taille 100



## Troisième cas : un échantillon de taille 100

On peut préciser les constats établis précédemment :

Par exemple, on peut calculer la probabilité d'obtenir d'un échantillon nous donnant une estimation inférieure à 10 %, 25 %, 40 %...

Ou une estimation comprise entre 40 % et 70 %...

Ou une estimation comprise entre 59 % et 63 %...

Ou une estimation supérieure à 80 %...

...

## Troisième cas : un échantillon de taille 100

On peut préciser les constats établis précédemment.  
Concrètement :

Probabilité que l'échantillon obtenu fournisse une estimation inférieure à 20 % =	0,0000000000000001072	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation inférieure à 40 % =	0,0007	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 40 et 80 % =	99,9964	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 45 et 75 % =	99,7682	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 50 et 70 % =	95,6328	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 55 et 65 % =	70,0965	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 59 et 63 % =	38,4109	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation égale à 61 % =	7,9867	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation supérieure à 80 % =	0,0085	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation supérieure à 90 % =	0,000000027231041	chances sur 100

On peut lire ces résultats de la manière suivante :

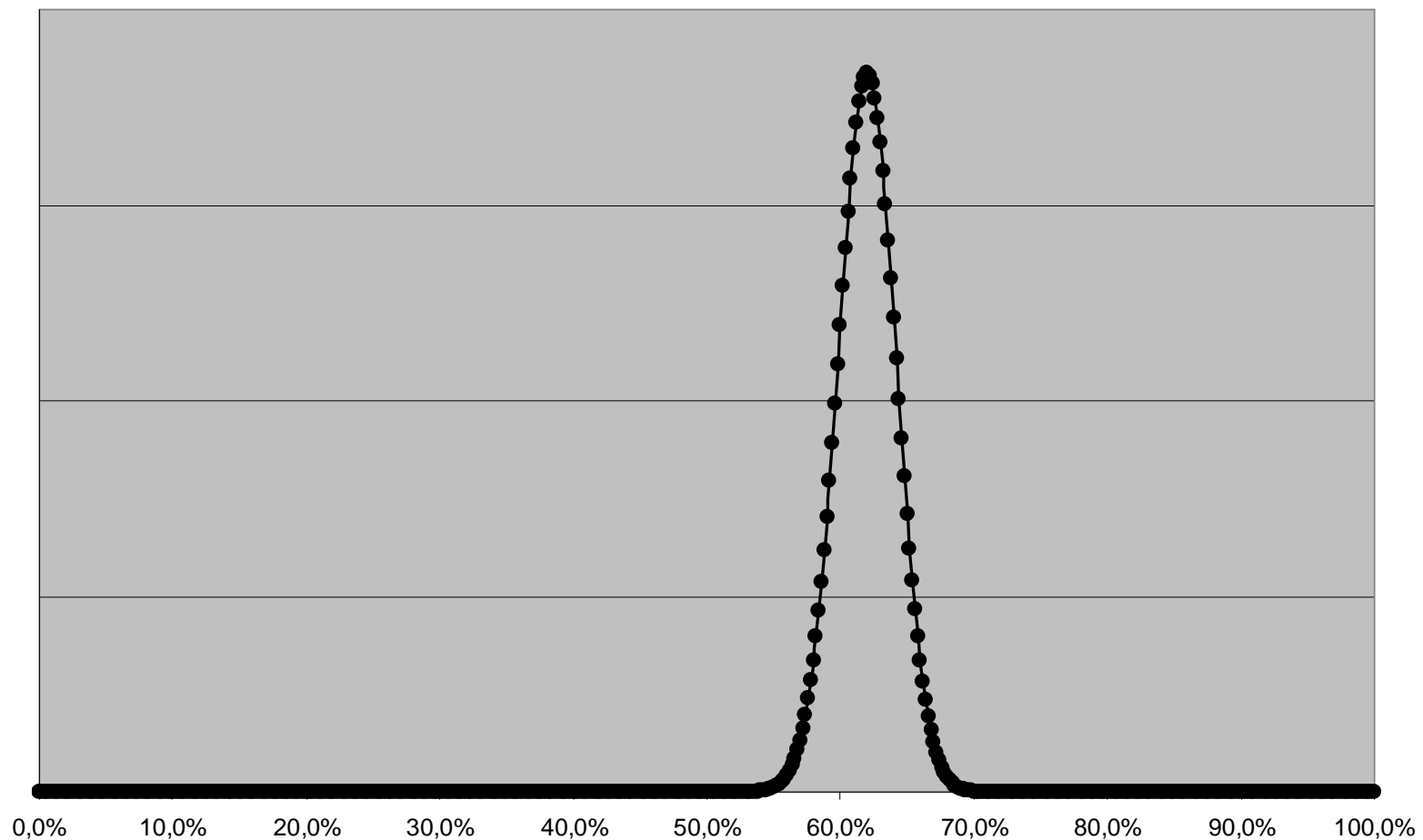
si 61 % des individus d'une population sont en position de répondre « oui »,  
et si nous disposons d'un échantillon de 100 individus de cette population,  
alors on a 95,6 % de chance que l'échantillon fournisse une estimation  
comprise entre 50 et 70 %...

## Quatrième cas : un échantillon de taille 500

- On peut simuler...
- On peut également calculer (calcul de probabilité – dénombrement)
  - quelle est la probabilité d'obtenir un échantillon contenant aucun « Oui » ?
  - quelle est la probabilité d'obtenir un échantillon contenant un seul « Oui » ?
  - ...
    - quelle est la probabilité d'obtenir un échantillon contenant 304 « Oui » ?
  - quelle est la probabilité d'obtenir un échantillon contenant 305 « Oui » ?
  - quelle est la probabilité d'obtenir un échantillon contenant 306 « Oui » ?
  - ...
    - quelle est la probabilité d'obtenir un échantillon contenant 499 « Oui » ?
    - quelle est la probabilité d'obtenir un échantillon contenant 500 « Oui » ?

## Troisième cas : un échantillon de taille 500

Fréquence des échantillons de taille 500 selon la part de "Oui" qu'ils contiennent



## Quatrième cas : un échantillon de taille 500

On peut préciser les constats établis précédemment.

Concrètement :

Probabilité que l'échantillon obtenu fournisse une estimation inférieure à 20 % =	0,0000000000000000	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation inférieure à 40 % =	0,0000000000000000	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 40 et 80 % =	100,00000000000000	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 45 et 75 % =	99,999999974500400	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 50 et 70 % =	99,992460183978500	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 55 et 65 % =	92,337124531541900	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation comprise entre 59 et 63 % =	61,570906708630700	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation égale à 61 % =	3,289005187102780	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation supérieure à 80 % =	0,0000000000000000	chances sur 100
Probabilité que l'échantillon obtenu fournisse une estimation supérieure à 90 % =	0,0000000000000000	chances sur 100

On peut lire ces résultats de la manière suivante :

si 61 % des individus d'une population sont en position de répondre « oui », et si nous disposons d'un échantillon de 500 individus de cette population, alors on a 99,9 % de chance que l'échantillon fournisse une estimation comprise entre 50 et 70 %...

## Constats

*a) Plus la taille de l'échantillon est grande, plus on a de chance d'avoir un échantillon qui donne une estimation assez proche de la vraie valeur...*

*b) On est capable de calculer la probabilité d'obtenir un échantillon encadrant la « vraie valeur » avec une précision donnée*

***DONC***

*on est capable d'obtenir un encadrement de la vraie valeur*

## Constat

*Ici encore, on a de « grandes chances »*

*d'avoir un échantillon qui donne*

*une estimation assez proche de la vraie valeur...*

## Reformulation !

la valeur estimée a de grande chance d'être dans un intervalle de valeurs contenant la vraie valeur

OU

la vraie valeur a de grande chance d'être dans un intervalle de valeurs contenant la valeur estimée

## En termes mathématiques :

$$\text{vraie valeur} - \varepsilon < \text{Estimation} < \text{vraie valeur} + \varepsilon$$

Donc :

$$\text{Estimation} - \varepsilon < \text{vraie valeur} < \text{Estimation} + \varepsilon$$

...

*on a de « grandes chances »*

*que la vraie valeur se situe*

*dans un intervalle autour de l'estimation...*

Intervalle de confiance

Niveau de confiance

## Terminologie :

*Intervalle de confiance*

*Niveau de confiance*

## Définition :

*L'intervalle de confiance est un intervalle de valeur qui contient la vraie valeur avec une probabilité égale au niveau de confiance*

## Questions

- a) *Si le niveau de confiance augmente, comment l'intervalle de confiance évolue-t-il ?*
  
- b) *Quel est le seul intervalle de confiance dont le niveau de confiance est 100 % ?*
  
- c) *Quels niveaux de confiance prendre en sciences sociales ?*

On peut calculer l'intervalle de confiance d'une valeur :

$$p - 1,96 \times \sqrt{\frac{p \times (1 - p)}{n}} \leq \text{vraie valeur} \leq p + 1,96 \times \sqrt{\frac{p \times (1 - p)}{n}}$$

En simplifiant (approximation) :

$$p - \frac{1}{\sqrt{n}} \leq \text{vraie valeur} \leq p + \frac{1}{\sqrt{n}}$$

## Exemple de calcul :

échantillon de taille **500**, valeur estimée = 35 %

32,9 % < Vraie valeur < 37,1 % (avec 95 % de chance)

Et avec la formule simplifiée :

30,5 % < Vraie valeur < 39,5 % (avec 95 % de chance)

## Exemple de calcul :

échantillon de taille **1000**, valeur estimée = 35 %

33,5 % < Vraie valeur < 36,5 % (avec 95 % de chance)

Et avec la formule simplifiée :

31,8 % < Vraie valeur < 38,1 % (avec 95 % de chance)

**Attention,  
l'intervalle de confiance  
d'une valeur ne permet  
pas de comparer  
deux pourcentages !**

**Deux pourcentages estimés :**

**Comment les comparer sachant que chacun d'entre eux sont des estimations, associées à des intervalles de confiance ?**

On peut calculer l'intervalle de confiance de deux valeurs

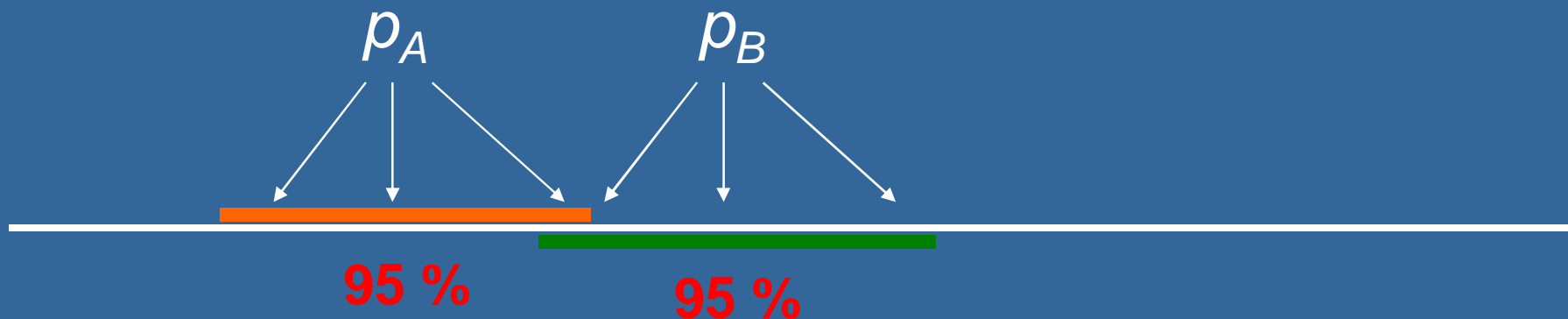


On aimerait comparer ces deux intervalles de confiance...

mais les intervalles de confiance sont des notions probabilistes !

## Conséquences de cette propriété ?

*Considérons par exemple :*



*Les deux IC se chevauchent...*

*Cela signifie-t-il que les deux pourcentages ne sont pas réellement différents ?*

*... indécidable ...*

**Pour comparer deux estimations, il faut :**

**1) Calculer l'intervalle de confiance de leur différence**

**2) Voir si cet intervalle contient la valeur 0**

**Si OUI**

- 0 est une valeur probable pour la différence
- Les deux estimations ne sont pas significativement différentes
- On accepte l'hypothèse qu'elles sont égales

**Si NON**

- 0 n'est pas une valeur probable pour la différence
- Les deux estimations sont significativement différentes
- On accepte l'hypothèse qu'elles ne sont pas égales

L'intervalle de confiance d'une différence :

*pour un niveau de confiance de 95 %*

$$(p_A - p_B) - 1,96 \times \sqrt{\frac{p_A \times (1 - p_A)}{n_A} + \frac{p_B \times (1 - p_B)}{n_B}} \leq \text{vraie valeur} \leq \dots$$
$$\dots (p_A - p_B) + 1,96 \times \sqrt{\frac{p_A \times (1 - p_A)}{n_A} + \frac{p_B \times (1 - p_B)}{n_B}}$$

On peut simplifier (au prix d'une approximation)

*pour un niveau de confiance de 95 %*

$$(p_A - p_B) - \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \leq \text{vraie valeur} \leq (p_A - p_B) + \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

## Exemples d'application

$$\left. \begin{array}{l} p_H = 28 \% ; n_H = 459 \\ p_F = 39 \% ; n_F = 236 \end{array} \right\} \rightarrow \text{IC (95 \%)} = [-18,4 \% ; -3 \%] \dots$$

$$\left. \begin{array}{l} p_H = 28 \% ; n_H = 459 \\ p_F = 33 \% ; n_F = 236 \end{array} \right\} \rightarrow \text{IC (95 \%)} = [-12,3 \% ; +2,2 \%] \dots$$

## Quelques repères

Écart minimum nécessaire pour juger significative  
une différence entre deux pourcentages  
(aux niveaux de confiance de 95 et 90 %)

Tailles des sous groupes	100	200	300	500	750	1000	2000
100	<u>13,5 %</u> (11,5 %)	<u>12,0 %</u> (10 %)	<u>11,0 %</u> (9,5 %)	<u>10,5 %</u> (9,0 %)	<u>10,0 %</u> (8,5 %)	<u>10,0 %</u> (8,5 %)	<u>10,0 %</u> (8,5 %)
200		<u>9,5 %</u> (8,0 %)	<u>9,0 %</u> (7,5 %)	<u>8,0 %</u> (7,0 %)	<u>7,5 %</u> (6,5 %)	<u>7,5 %</u> (6,5 %)	<u>7,0 %</u> (6,0 %)
300			<u>8,0 %</u> (6,5 %)	<u>7,0 %</u> (6,0 %)	<u>6,5 %</u> (5,5 %)	<u>6,5 %</u> (5,5 %)	<u>6,0 %</u> (5,0 %)
500				<u>6,0 %</u> (5,0 %)	<u>5,5 %</u> (4,5 %)	<u>5,5 %</u> (4,5 %)	<u>5,0 %</u> (4,0 %)
750					<u>5,0 %</u> (4,0 %)	<u>4,5 %</u> (4,0 %)	<u>4,0 %</u> (3,5 %)
1000						<u>4,5 %</u> (3,5 %)	<u>3,5 %</u> (3,0 %)
2000							<u>3,0 %</u> (2,5 %)

L2...

## En pratique

- On ne calcule pas systématiquement les IC des pourcentages.
- On peut établir les ordres de grandeur des incertitudes sur les pourcentages pour l'échantillon considéré.
- Il faut connaître des ordres de grandeur
  
- Et si on ne travaille pas sur un échantillon aléatoire ?
  - Les calculs des IC fournissent malgré tout des ordres de grandeur des erreurs « mécaniques » (auxquelles on peut ajouter d'autres sources d'erreur et de biais...)