Introduction aux traitements statistiques d'enquêtes sociologiques

Cours 2 : Principe de l'inférence

Damien Cartron et Martin Chevalier

Année universitaire 2020-2021

Rappel des épisodes précédents

Calcul de **statistiques descriptives univariées** :

- statistiques adaptées aux **variables qualitatives** : effectifs, pourcentages ;
- statistiques de **tendance centrale** : moyenne, médiane, etc. ;
- statistiques de **dispersion** : variance empirique, écart-type, quantiles, etc. ;
- indicateurs d'**inégalité** : courbe de Lorenz, indicateur de Gini.

Le problème du jour

Jusqu'à présent, les statistiques descriptives ont été calculées sur des jeux de données sans s'interroger sur leur **représentativité**.

Or en pratique en sciences sociales, la quasi-totalité des données quantitatives proviennent d'**enquêtes par sondage** : au lieu d'interroger l'ensemble de la **population d'intérêt**, l'enquête ne porte que sur un **échantillon**.

Exemple L'enquête Emploi en continu (EEC) permet de calculer chaque trimestre le taux de chômage des 15 ans et plus (53,5 M de personnes en 2018) à partir d'un échantillon d'environ 110 000 personnes.

Question Le fait que les données proviennent d'enquêtes par sondage est-il de nature à **modifier la manière d'interpréter les statistiques descriptives** ?

Objectifs de la séance

1. Comprendre l'impact du sondage sur le calcul et l'interprétation de statistiques descriptives
2. Acquérir des points de repères en matière d'inférence statistique
3. Savoir construire l'intervalle de confiance d'une moyenne
4. Savoir utiliser les pondérations d'une enquête statistique par sondage

Statistiques descriptives et enquête par sondage

Les enquêtes par sondage en sciences sociales

Les enquêtes par sondage constituent une **part importante** des données quantitatives en sciences sociales.

Pour ce qui concerne les enquêtes de la statistique publique française, leur échantillon est **tiré aléatoirement** au sein d'une base de données couvrant l'ensemble de la population (la **base de sondage**).

En pratique, la manière de tirer l'échantillon, qu'on appelle **plan de sondage**, peut varier grandement d'une enquête à une autre (*cf.* séance 8)

Une enquête par sondage est caractérisée entre autre par son **taux de sondage** souvent noté f :

$$f = \frac{\text{Nombre d'unités dans l'échantillon}}{\text{Nombre d'unités dans la population}} = \frac{n}{N}$$

Les enquêtes par sondage en sciences sociales (2)

Exemple Le taux de sondage de l'EEC est de l'ordre de

$$f = \frac{110\ 000}{53\ 500\ 000} = 0,2\ \%$$

Les taux de sondage des enquêtes par sondage sont en général **extrêmement faibles** :

- sondage électoral : $n = 1\ 000$, $N = 50$ M, $f = 0,002\ \%$
- enquête standard de l'Insee : $n = 20\ 000$, $N = 65$ M, $f = 0,03\ \%$
- enquête Epicov : $n = 370\ 000$, $N = 53$ M, $f = 0,7\ \%$

Exception notable Le recensement de la population : exhaustif sur les communes de moins de 10 000 habitants, taux de sondage de 40 % sur les communes de 10 000 habitants ou plus.

Impact du sondage sur les statistiques descriptives

Que les données soient issues d'un sondage introduit une **incertitude** dans la valeur des statistiques descriptives :

- si les individus avec un **salaire faible** sont sur-représentés dans l'échantillon, alors le salaire moyen sera **sous-estimé** par l'enquête ;
- mais si au contraire ce sont les individus avec un **salaire élevé** qui sont sur-représentés dans l'échantillon, alors le salaire moyen sera **sur-estimé**.

L'estimation du salaire moyen à partir de l'enquête n'est donc **pas parfaite**, elle est entâchée d'une **marge d'erreur**. Pour autant l'enquête est bien **représentative**.

Remarque On parle en général d'« **estimateurs** » (par opposition à la « **vraie valeur** » ou à la **valeur dans la population**) pour souligner cette incertitude.

Une petite simulation vaut mieux qu'un long discours

Il est souvent très utile en sondage de faire des simulations :

- pour vérifier un résultat théorique ;
- pour confirmer une intuition ;
- pour expliquer un phénomène complexe.

En pratique, cela revient à :

1. Générer des données (aléatoires) sur la population
2. Simuler le tirage de l'échantillon un très grand nombre de fois
3. Comparer la valeur de l'estimateur (ici le salaire moyen) avec la vraie valeur dans la population

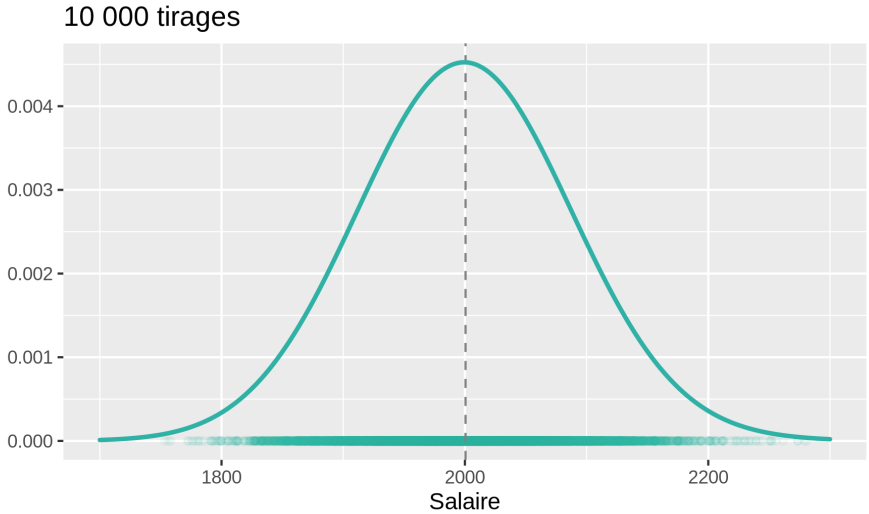
Une petite simulation vaut mieux qu'un long discours

En l'espèce ici :

- population de 25 M d'actifs occupés ;
- salaire simulé avec une moyenne de 2 000 euros ;
- tirage aléatoire de 10 000 échantillons de taille 500 individus ;
- calcul de la moyenne empirique des salaires des individus sélectionnés.

Représentation des **estimations du salaire moyen pour les différents échantillons**.

Une petite simulation vaut mieux qu'un long discours



Une petite simulation vaut mieux qu'un long discours

Plusieurs remarques :

- la **variabilité des estimations d'un échantillon à l'autre est réelle** : certaines s'écartent de la vraie valeur de plus de 200 euros, soit 10 % de la vraie valeur ;
- néanmoins, **en moyenne sur l'ensemble des échantillons tirés on retrouve la vraie valeur**, et ce d'autant plus que le nombre d'échantillons est important ;
- plus encore, quand le nombre de simulations augmente on perçoit une régularité dans la distribution des estimations.

Les propriétés d'un estimateur par sondage

Ces différents éléments correspondent aux propriétés classique d'un estimateur dans une enquête par sondage :

- il est en général **sans biais** : si on pouvait calculer la moyenne des estimations sur tous les échantillons possibles, on retrouverait la vraie valeur ;
- il présente une **variance plus ou moins importante** ;
- il suit une **distribution théorique** qui rend cette variance calculable $\rightarrow$ c'est tout l'enjeu de l'**inférence** (*cf.* partie suivante).

Variante 1 : avec une variable d'intérêt plus dispersée

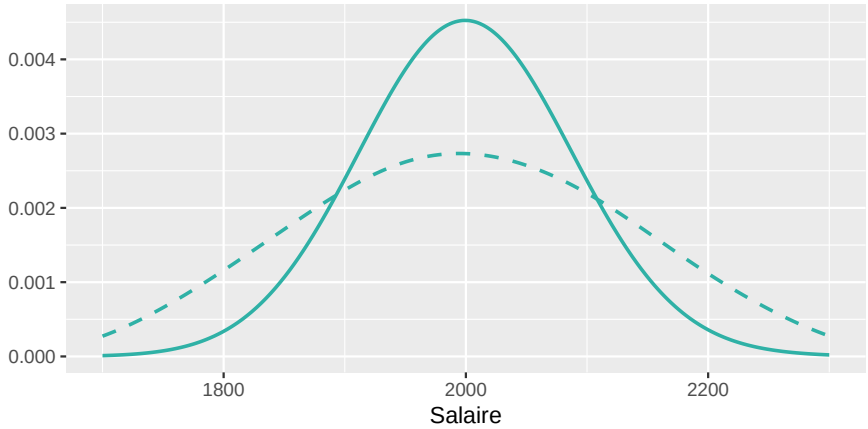
L'intérêt de procéder par simulation est de pouvoir **très facilement modifier** des paramètres qui sont habituellement fixés.

En particulier, il est possible dans ce cadre de **faire varier à volonté la dispersion** de la variable d'intérêt.

En l'espèce ici, on simule **un scénario alternatif** où l'écart-type de la variable d'intérêt est environ **deux fois plus élevé** que dans le scénario d'origine.

Moralité Plus la variable d'intérêt est **dispersée**, plus l'estimateur de sa moyenne est **imprécis**.

Variante 1 : avec une variable d'intérêt plus dispersée



Variante 2 : avec un échantillon plus grand

La taille de l'échantillon a un **impact déterminant** sur l'imprécision des estimateurs construits à partir de l'enquête.

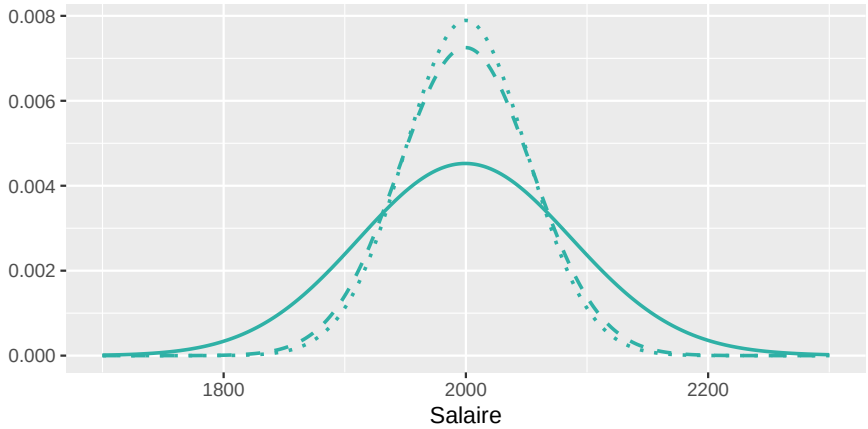
Plus un échantillon est grand, plus les estimateurs sont précis
$\rightarrow$ moins de chances que l'échantillon contienne **uniquement le même « profil »** d'individus.

On simule à nouveau **deux scénarios alternatifs** :

- échantillon **dix fois plus grand** que dans le scénario d'origine ;
- échantillon **cent fois plus grand** que dans le scénario d'origine.

Moralité Plus l'échantillon est **grand**, plus l'estimateur de la moyenne est **précis** (pour une variable d'intérêt donnée).

Variante 2 : avec un échantillon plus grand



L'inférence : des outils pour tenir compte de l'incertitude

Principe de l'inférence

Dans un travail de recherche en sciences sociales cependant, on n'est **précisément pas** dans un cadre de simulations : on dispose toujours d'**un seul et unique échantillon** !

→ **Comment tenir compte de la variabilité associée au sondage ?**

C'est là qu'intervient l'**inférence**, *ie* un ensemble d'outils statistiques visant à élaborer un **discours scientifique** en présence d'**incertitude**.

En pratique : on va pouvoir s'appuyer sur **des concepts et des résultats statistiques** pour évaluer la **variabilité d'un estimateur à partir du seul échantillon disponible**.

Notion de variable aléatoire

Une variable aléatoire est une **fonction mathématique** qui associe des **valeurs numériques** à un **ensemble d'éventualités**.

Exemple Jet d'un dé parfaitement équilibré

L'ensemble des éventualités est : $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ et la probabilité de chacune est $1/6$. Dans ce contexte, on peut assez naturellement définir la variable aléatoire X :

$$X : \begin{array}{ccc} \Omega & \to & \{1,2,3,4,5,6\} \\ \omega & \mapsto & \begin{cases} 1 \text{ si } \boxdot, 2 \text{ si } \boxdot, 3 \text{ si } \boxdot, \\ 4 \text{ si } \boxdot, 5 \text{ si } \boxdot, 6 \text{ si } \boxdot \end{cases} \end{array}$$

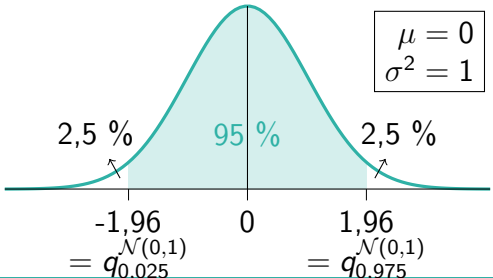
On sait alors que la variable aléatoire X suit une **loi de probabilité uniforme** sur $\{1,2,3,4,5,6\}$.

Notion de loi de probabilités

Les lois de probabilités constituent des **éléments-clés de la théorie statistique**.

Elles permettent de **décrire le comportement de variables aléatoires** et ainsi de les distinguer les unes des autres.

Exemple Loi normale (ou gaussienne) notée $\mathcal{N}(\mu, \sigma^2)$



Notion de loi de probabilités

En pratique, on dira par exemple qu'une variable aléatoire X « **suit** » une loi normale centrée réduite et on notera

$$X \hookrightarrow \mathcal{N}(0,1)$$

De la sorte, on a une information sur le comportement de la variable aléatoire X, en particulier sur la **fréquence à laquelle elle est susceptible de prendre certaines valeurs** :

- $\mathbb{P}[X > 1,96] = 2,5\ \%$
- $\mathbb{P}[X < -1,96] = 2,5\ \%$
- $\mathbb{P}[-1,96 < X < 1,96] = 95\ \%$

En bref, quand une variable aléatoire suit une certaine loi, on sait **avec quelle probabilité elle prend certaines valeurs**.

Quantiles de la loi normale centrée réduite

En particulier pour la loi normale centrée réduite, on a les **quantiles suivants** :

γ	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
$q_\gamma^{\mathcal{N}(0,1)}$	-2,58	-2,33	-1,96	-1,64	-1,28	1,28	1,64	1,96	2,33	2,58

Autrement dit on sait ainsi qu'une variable aléatoire suivant une loi normale centrée réduite :

- a très peu de chances (moins de 0,5 %) de prendre une valeur inférieure à -2,58 ;
- en revanche, elle a 80 % de chances de prendre une valeur dans l'intervalle [-1,28 ; 1,28].

C'est précisément ce type de résultat qui permet de **déterminer la plage de variation attendue** pour un estimateur issu d'un sondage.

Construction de l'intervalle de confiance d'une moyenne

Théorème central limite dérivé pour les sondages

Dans le cadre d'un **sondage aléatoire simple** (*cf.* séance 8), on peut montrer que l'estimateur de la moyenne d'une variable Y (noté $\hat{\bar{Y}}$) suit une loi normale :

$$\hat{\bar{Y}} \hookrightarrow \mathcal{N}(\bar{Y}, \hat{V}(\hat{\bar{Y}}))$$

où $\bar{Y}$ est la vraie valeur de la moyenne (dans la population) et $\hat{V}(\hat{\bar{Y}})$ l'**estimateur de la variance** de $\hat{\bar{Y}}$:

$$\hat{V}(\hat{\bar{Y}}) = (1 - f)\frac{s_Y^2}{n}$$

avec f le taux de sondage, s_Y^2 la variance empirique de Y dans l'échantillon et n le nombre d'observations dans l'échantillon.

Théorème central limite dérivé pour les sondages

Remarque Il s'agit d'une dérivation dans le domaine des sondages d'un théorème de portée plus générale, le **théorème central limite**.

En substance, le théorème central limite énonce qu'en calculant la moyenne de variables aléatoires toutes de même loi et indépendantes les unes des autres, on obtient une **nouvelle variable aléatoire qui suit une loi normale** et dont les paramètres sont connus.

Ce théorème fondamental de la statistique est sous-jacent à de nombreux résultats utilisés en inférence, notamment de nombreux **tests statistiques** (*cf.* séance 4).

Construction d'un intervalle de confiance

On tire du résultat précédent que :

$$\frac{\hat{\bar{Y}} - \bar{Y}}{\sqrt{\hat{V}(\hat{\bar{Y}})}} \hookrightarrow \mathcal{N}(0,1)$$

À partir de là, il est possible de déterminer un **intervalle** dans lequel la probabilité que se trouve la vraie valeur de la moyenne est **maîtrisée**.

On va pour ce faire s'appuyer sur ce qu'on sait du **comportement de la loi normale centrée réduite**.

Construction d'un intervalle de confiance à 95 %

D'après les quantiles de la loi normale centrée réduite, on sait que

$$\begin{aligned} 95\ \% &= \mathbb{P}\left[-1,96 \leq \frac{\hat{\bar{Y}} - \bar{Y}}{\sqrt{\hat{V}(\hat{\bar{Y}})}} \leq 1,96\right] \\ &= \mathbb{P}\left[-1,96\sqrt{\hat{V}(\hat{\bar{Y}})} \leq \hat{\bar{Y}} - \bar{Y} \leq 1,96\sqrt{\hat{V}(\hat{\bar{Y}})}\right] \\ &= \mathbb{P}\left[\hat{\bar{Y}} - 1,96\sqrt{\hat{V}(\hat{\bar{Y}})} \leq \bar{Y} \leq \hat{\bar{Y}} + 1,96\sqrt{\hat{V}(\hat{\bar{Y}})}\right] \end{aligned}$$

Autrement dit, on en déduit que **l'intervalle suivant a 95 % de chance de contenir la vraie valeur** :

$$\hat{IC}_{95\ \%}(\hat{\bar{Y}}) = \left[\hat{\bar{Y}} - 1,96\sqrt{\hat{V}(\hat{\bar{Y}})}; \hat{\bar{Y}} + 1,96\sqrt{\hat{V}(\hat{\bar{Y}})}\right]$$

Il est possible de réécrire ce qui précède sous la forme :

$$\hat{IC}_{95\ \%}(\hat{\bar{Y}}) = \hat{\bar{Y}} \pm \underbrace{1,96\sqrt{\hat{V}(\hat{\bar{Y}})}}_{DL_{95\ \%}(\hat{\bar{Y}})}$$

où $DL_{95\ \%}(\hat{\bar{Y}})$ désigne la **demi-longueur de l'intervalle de confiance**.

Cette réécriture permet de mettre en évidence **plusieurs éléments** :

- l'intervalle de confiance est **symétrique** autour de la valeur estimée dans l'échantillon $\hat{\bar{Y}}$;
- plus la demi-longueur de l'intervalle de confiance est **grande**, plus l'estimateur est **imprécis**.

En développant les formules :

$$DL_{95\ \%}(\hat{\bar{Y}}) = 1,96 \times \sqrt{(1-f)\frac{s_Y^2}{n}}$$

En règle générale $f = \frac{n}{N} \approx 0$ aussi on peut retenir :

$$DL_{95\ \%}(\hat{\bar{Y}}) \approx 1,96 \times \frac{s_Y}{\sqrt{n}}$$

En d'autres termes :

- plus l'**écart-type de la variable d'intérêt** s_Y est **élevé**, plus l'intervalle de confiance est **étendu** $\rightarrow$ **estimateur imprécis** ;
- plus la **taille de l'échantillon** n est **élevée**, plus l'intervalle de confiance est **resserré** $\rightarrow$ **estimateur précis**.

Choisir le niveau de l'intervalle de confiance

Dans les formules précédentes, la valeur 1,96 est directement liée au **niveau de l'intervalle de confiance**, c'est-à-dire qu'il soit « à 95 % ».

En réalité il est possible de construire des intervalles de confiance à **n'importe quel niveau**.

Exemple Intervalle de confiance à 99 %

- On repart du résultat fondamental :

$$\frac{\hat{\bar{Y}} - \bar{Y}}{\sqrt{\hat{V}(\hat{\bar{Y}})}} \hookrightarrow \mathcal{N}(0,1)$$

  mais on l'applique en utilisant **d'autres quantiles** de la loi normale centrée réduite.
- On lit en particulier qu'il faut prendre l'intervalle **[-2,58 ; 2,58]** pour capter **99 % des valeurs de la distribution**.

Choisir le niveau de l'intervalle de confiance

On en tire que :

$$
\begin{aligned}
99\ \% &= \mathbb{P}\left[-2,58 \leq \frac{\hat{\bar{Y}} - \bar{Y}}{\sqrt{\hat{V}(\hat{\bar{Y}})}} \leq 2,58\right] \\
&= \mathbb{P}\left[-2,58\sqrt{\hat{V}(\hat{\bar{Y}})} \leq \hat{\bar{Y}} - \bar{Y} \leq 2,58\sqrt{\hat{V}(\hat{\bar{Y}})}\right] \\
&= \mathbb{P}\left[\hat{\bar{Y}} - 2,58\sqrt{\hat{V}(\hat{\bar{Y}})} \leq \bar{Y} \leq \hat{\bar{Y}} + 2,58\sqrt{\hat{V}(\hat{\bar{Y}})}\right]
\end{aligned}
$$

et ainsi

$$
\hat{IC}_{99\ \%}(\hat{\bar{Y}}) = \left[\hat{\bar{Y}} - 2,58\sqrt{\hat{V}(\hat{\bar{Y}})}; \hat{\bar{Y}} + 2,58\sqrt{\hat{V}(\hat{\bar{Y}})}\right] = \hat{\bar{Y}} \pm 2,58\sqrt{\hat{V}(\hat{\bar{Y}})}
$$

Bien interpréter le niveau de l'intervalle de confiance

Le niveau de l'intervalle de confiance est en rapport direct avec le **risque de se tromper** en affirmant que la **vraie valeur se trouve dans l'intervalle**.

Exemple Si l'IC à 95 % du salaire moyen pour un échantillon est [1 610 ; 1 990], alors :

- la probabilité que cet intervalle contienne la vraie valeur du salaire moyen est de 95 %... :
- ...autrement dit il y a un risque de 5 % de se tromper en affirmant que le salaire moyen dans la population est compris entre 1 610 et 1 990 euros.

En choisissant un **niveau plus élevé** (par exemple 99 %) :

- on augmente mécaniquement sa largeur (1,96 remplacé par 2,58 dans les formules)... ;
- ...et on diminue ainsi le risque de « passer à côté » de la vraie valeur $\rightarrow$ celui-ci n'est plus que de 1 %.

Si un intervalle de confiance est de 95 %, alors cela signifie qu'**il est censé contenir la vraie valeur dans 95 % des cas**.

Il est possible de **vérifier cette propriété par simulation** :

1. Dans une population connue, on tire un grand nombre d'échantillons
2. Pour chaque échantillon, on calcule l'intervalle de confiance de l'estimateur et on détermine s'il contient la vraie valeur
3. Cela doit être le cas dans environ 95 % des échantillons

On procède à cette vérification pour les simulations de la première partie : l'intervalle de confiance estimé contient la vraie valeur dans **9 497 échantillons sur 10 000**.

Construction de l'intervalle de confiance d'une proportion

Construction de l'intervalle de confiance d'une proportion
La proportion est un cas particulier de moyenne

Quand un traitement porte sur une variable qualitative, on ne manipule pas des moyennes mais des **proportions** (*ie* des pourcentages).

Néanmoins, on peut voir une proportion comme un **cas particulier de moyenne**.

Exemple Il est équivalent de :

- calculer le pourcentage d'inactifs à partir de la variable qualitative « Position sur le marché du travail » (Actif occupé, chômeur, inactif) ;
- calculer la moyenne de la **variable indicatrice** qui vaut 1 quand l'individu est inactif et 0 sinon.

On peut donc **appliquer directement** les résultats vus précédemment au cas d'une proportion.

En particulier, on a défini :

$$\hat{V}(\hat{\bar{Y}}) = (1-f)\frac{s_Y^2}{n}$$

En notant $\hat{p}$ l'estimateur de la proportion (comprise entre 0 et 1), on peut réécrire cette variance :

$$\hat{V}(\hat{p}) = (1-f)\frac{s_p^2}{n}$$

Or on peut montrer que $s_p^2 = \frac{\hat{p}(1-\hat{p})n}{n-1}$

On obtient ainsi dans le cas d'une proportion $\hat{V}(\hat{p}) = (1-f)\frac{\hat{p}(1-\hat{p})}{n-1}$

Application : sondages électoraux

On assimile en général les sondages électoraux à des sondages aléatoire simples : les résultats vus précédemment **s'appliquent donc à eux**.

En considérant le **taux de sondage comme négligeable**, on peut ainsi très facilement évaluer l'intervalle de confiance de leurs résultats avec :

$$IC_{95\ \%}(\hat{p}) = \hat{p} \pm 1,96\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$$

Remarques

- Cette formule est valable **quelle que soit la taille de la population d'intérêt** N : celle-ci n'intervient que par l'intermédiaire du taux de sondage f en pratique négligeable pour les sondages électoraux.
- Cela illustre à quel point **la taille de l'échantillon est beaucoup plus déterminante** en matière de précision que le taux de sondage.

Construction de l'intervalle de confiance d'une proportion

Exemple : Premier tour des élections présidentielles de 2002

	n	$\hat{p}$	$IC_{95\ \%}(\hat{p})$	Score le 21 avril 2002
Chirac	989	20,0 %	[17,5 % ; 22,5 %]	19,9 %
Jospin	989	18,0 %	[15,6 % ; 20,4 %]	16,2 %
Le Pen	989	14,0 %	[11,8 % ; 16,2 %]	16,9 %

Remarques

1. Les intervalles de confiance de Chirac et Jospin **contiennent la vraie valeur**, mais pas celui de Le Pen ;
2. Au-delà des 5 % d'erreur statistique, **d'autres biais sont possibles** : seuls 59 % des sondés savaient avec certitude pour qui ils allaient voter et 19 % de ceux-ci ont refusé de dire pour qui ;
3. La demi-longueur de l'intervalle de confiance **décroît avec** $\hat{p}$: 2,5 pts pour Chirac, 2,4 pts pour Jospin, 2,2 pts pour Le Pen.

Interprétation de la demi-longueur de l'IC dans le cas d'une proportion

La quantité $p(1-p)$ est maximale sur $[0\,;\,1]$ pour $p = 0,5$.

Pour une **même taille d'échantillon**, la demi-longueur de l'IC est d'**autant plus grande que la proportion à estimer est proche de 50 %**.

	5 %	**10 %**	**25 %**	**50 %**	**75 %**	**90 %**	**95 %**
1 000	1,4 pts	1,9 pts	2,7 pts	3,1 pts	2,7 pts	1,9 pts	1,4 pts
2 000	1,0 pts	1,3 pts	1,9 pts	2,2 pts	1,9 pts	1,3 pts	1,0 pts
5 000	0,6 pts	0,8 pts	1,2 pts	1,4 pts	1,2 pts	0,8 pts	0,6 pts
10 000	0,4 pts	0,6 pts	0,8 pts	1,0 pts	0,8 pts	0,6 pts	0,4 pts

À retenir Quand la proportion à estimer est **proche de 50 %** et que l'échantillon est de **taille 1 000**, l'incertitude est **de l'ordre de $\pm$ 3 pts**.

Les pondérations : des intermédiaires de calcul indispensables

Définition et caractéristiques des pondérations

En pratique dans un fichier d'enquête, les caractéristiques essentielles du sondage sont incorporées sous la forme d'une **variable de pondération**.

Concrètement, c'est cette variable qui permet de **passer de l'échantillon de l'enquête à la population d'intérêt**.

Le « poids » de chaque individu de l'échantillon peut être interprété comme le **nombre d'individus** de la population d'intérêt **qu'il représente**.

En règle générale, on fait en sorte que **la somme des poids des individus de l'échantillon coïncide exactement avec la taille de la population d'intérêt**.

Pondérations dans l'enquête Emploi en continu

cours.R* ee12t4

Filter

	IDENT IDENT	NOI NOI	SEXE SEXE	AGE AGE	SALRED SALRED	EXTRI1613
1	G0A56JP6	01	2	66	596	1776.563
2	G0A56JP6	02	1	29	*NA*	1776.563
3	G0A56JR6	01	2	27	2700	2044.920
4	G0A56JS6	01	2	29	2666	1897.519
5	G0A56JT6	01	1	31	11967	1754.300
6	G0A56JU6	01	2	24	*NA*	1643.808
7	G0A56JV6	01	1	27	2000	2191.007
8	G0A56JV6	02	2	25	2800	2191.007
9	G0A56JW6	01	1	35	2333	1616.748

Showing 1 to 10 of 34,913 entries, 6 total columns

EXTRI1613 est une des variables de pondération de l'EEC.

Pourquoi utiliser les pondérations ?

1. Pour estimer des totaux sur l'ensemble de la population d'intérêt et pas seulement sur l'échantillon.
   Exemple Population en emploi.

2. Pour corriger les sur- ou sous-représentation de l'échantillon : bien souvent les pondérations ne sont pas toutes égales, certains individus sont sur-représentés et d'autres sous-représentés.
   Exemple Enquête Patrimoine, enquêtes Santé.
   $\rightarrow$ Ne pas utiliser les pondérations conduit à **accorder trop d'importance aux individus sur-représentés** et pas assez aux autres.

3. Pour obtenir une **première évaluation de l'imprécision liée au sondage** : SAS et R comportent des fonctions qui, à partir du poids de sondage (entre autres), estiment des intervalles de confiance.

Comment utiliser les pondérations ?

1. SAS : instructions WEIGHT de les procédures classiques (FREQ, MEANS, UNIVARIATE) + option VARDEF = WGT ;
2. SAS : procédures SURVEYFREQ et SURVEYMEANS avec instruction WEIGHT + d'éventuelles précisions sur le sondage (strates, etc. → *cf.* séance 8) ;
3. R : *package* survey.

Remarque Sauf dans les sondages les plus simples, ces méthodes ne peuvent fournir qu'une **première approximation de l'imprécision**
→ des **méthodes d'estimation de la précision exactes existent**.