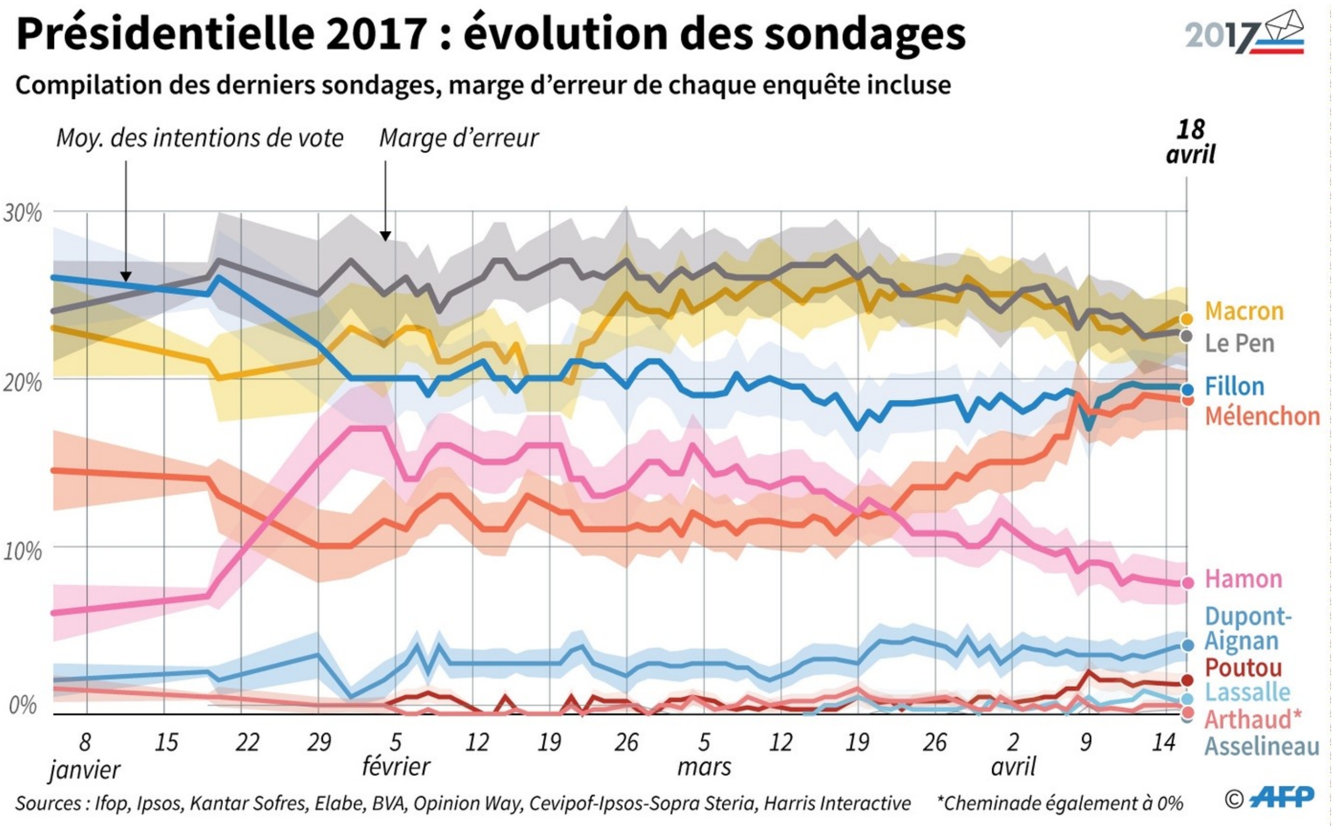
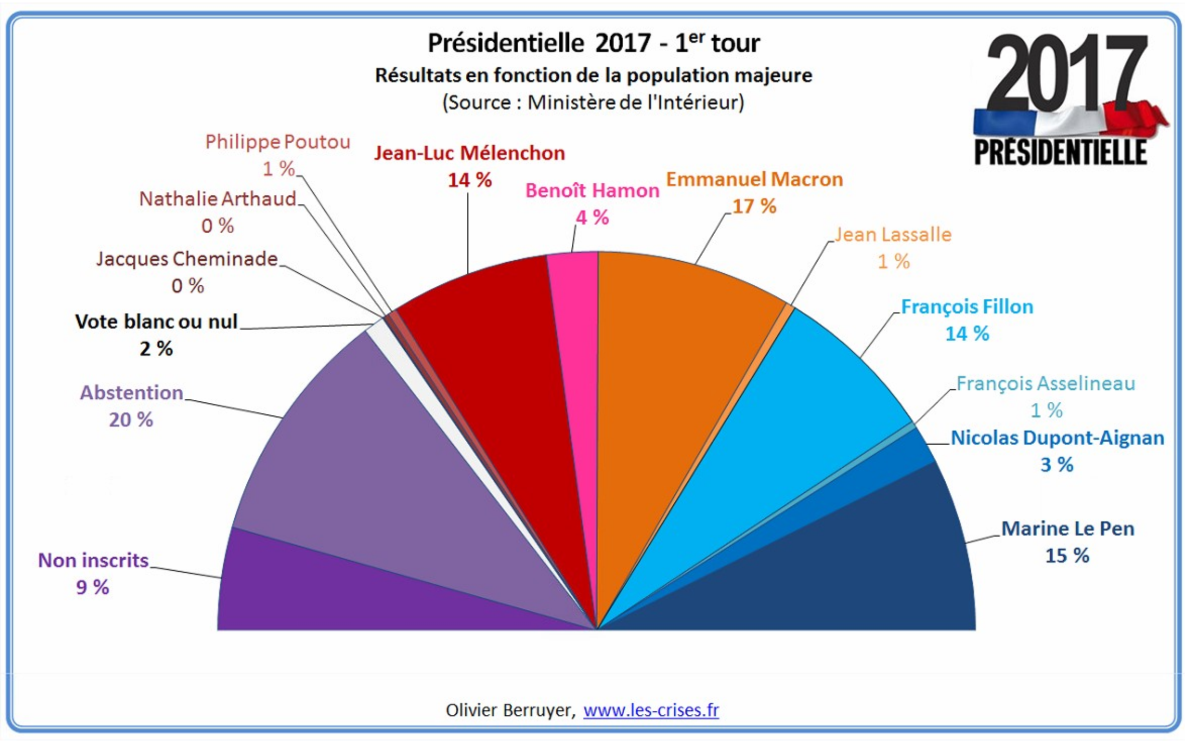
ESTIMATION : INTERVALLE(S) DE CONFIANCE





I. De la fluctuation d'échantillonnage à un intervalle aléatoire

PROPRIÉTÉ. Soit $X_n \sim \mathcal{B}(n\,;p)$. On note $F_n = \dfrac{X_n}{n}$.

Lorsque n est assez grand, l'intervalle $\left[F_n - \dfrac{1}{\sqrt{n}}\,;F_n + \dfrac{1}{\sqrt{n}}\right]$ contient la proportion p avec une probabilité <u>environ</u> égale à 0,95.

<u>Démonstration</u> : On a dès la Seconde que : $\mathrm{p}\left(F_n \in \left[p - \dfrac{1}{\sqrt{n}}\,;p + \dfrac{1}{\sqrt{n}}\right]\right) \approx 0{,}95$.

Or : $F_n \in \left[p - \dfrac{1}{\sqrt{n}}\,;p + \dfrac{1}{\sqrt{n}}\right] \Leftrightarrow$

Donc la probabilité pour que l'intervalle aléatoire $\left[F_n - \dfrac{1}{\sqrt{n}}\,;F_n + \dfrac{1}{\sqrt{n}}\right]$ contienne p est environ égale à 0,95.

Ce « lorsque n est assez grand » est peu précis. On peut dire, plus précisément :

ROC

PROPRIÉTÉ. Soit $X_n \sim \mathcal{B}(n\,;p)$. On note $F_n = \dfrac{X_n}{n}$.

Il existe un entier n_0 tel que, pour tout entier naturel $n \geqslant n_0$:

$$\mathrm{p}\left(F_n \in \left[p - \frac{1}{\sqrt{n}}\,;p + \frac{1}{\sqrt{n}}\right]\right) > 0{,}95 .$$

<u>Démonstration</u> :

On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$.

On admet qu'il existe un entier n_0 tel que, pour tout entier naturel $n \geqslant n_0$:

$$\mathrm{p}(-2 \leqslant Z_n \leqslant 2) > 0{,}95 \text{ où } T \sim N(0\,;1).$$

1. Démontrer que, pour tout entier naturel $n \geqslant n_0$:

$$\mathrm{p}\left(p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{X_n}{n} \leqslant p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) > 0{,}95 .$$

2. On définit sur $[0\,;1]$ la fonction f par $f(p) = p(1-p)$.

a) Étudier la fonction f et dresser son tableau de variations sur $[0\,;1]$.

b) En déduire que $0 \leqslant \sqrt{p(1-p)} \leqslant \dfrac{1}{2}$.

c) En déduire que, pour tout entier naturel $n \geqslant n_0$: $\mathrm{p}\left(p - \dfrac{1}{\sqrt{n}} \leqslant \dfrac{X_n}{n} \leqslant p + \dfrac{1}{\sqrt{n}}\right) > 0{,}95$.

Et en remarquant que $F_n \in \left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}}\right] \Leftrightarrow p \in \left[F_n-\frac{1}{\sqrt{n}} ; F_n+\frac{1}{\sqrt{n}}\right]$, on en déduit :

PROPRIÉTÉ. Soit $X_n \sim \mathcal{B}(n ; p)$. On note $F_n=\frac{X_n}{n}$.

Il existe un entier n_0 tel que, pour tout entier naturel $n \geqslant n_0$, l'intervalle $\left[F_n-\frac{1}{\sqrt{n}} ; F_n+\frac{1}{\sqrt{n}}\right]$ contient la proportion p avec une probabilité supérieure ou égale à 0,95.

Remarque : cet entier n_0 dépend de p, et peut varier considérablement.
A l'aide d'algorithmes, on peut obtenir le tableau suivant :

p	0,35	0,36	0,37	0,38	0,39	0,4	0,41	0,42	0,43	0,44	0,45	0,46	0,47	0,48	0,49	0,5
n_0	31	30	36	64	56	81	90	120	143	209	271	288	304	399	399	529

II. Un intervalle de confiance au niveau 0,95

On souhaite connaître, dans une population, la valeur d'une proportion p (proportion des gauchers en France, intentions de vote pour une élection, …).
Pour des raisons matérielles, financières ou autres[1], on ne peut pas toujours réunir les données concernant la population tout entière.
On va donc ***estimer la proportion p que l'on cherche à partir de la fréquence f observée dans un échantillon***.

Rappel : on sait que cette fréquence observée va varier d'un échantillon à l'autre, c'est la fluctuation d'échantillonnage autour de p.

Grâce aux propriétés précédentes, on a vu que :

Si n est assez grand, la probabilité qu'un échantillon de taille n de cette population ait une fréquence f qui appartienne à l'intervalle de fluctuation $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}}\right]$ est strictement supérieure à 95 %.
C'est la fluctuation d'échantillonnage.

Soit p une proportion inconnue dans une population.
Pour un échantillon de taille n de cette population, on calcule la fréquence f.
Alors la probabilité que l'intervalle $\left[f-\frac{1}{\sqrt{n}} ; f+\frac{1}{\sqrt{n}}\right]$ contienne p est strictement supérieure à 95 %.

Autrement dit, parmi tous les échantillons de taille n qu'on peut obtenir, au moins 95 % d'entre eux sont tels que l'intervalle $\left[f-\frac{1}{\sqrt{n}} ; f+\frac{1}{\sqrt{n}}\right]$ contient la proportion p.

On dit donc que $\left[f-\frac{1}{\sqrt{n}} ; f+\frac{1}{\sqrt{n}}\right]$ *est **un intervalle de confiance** de la proportion inconnue* p *à un niveau de confiance de 0,95.*

1 par exemple, on ne peut pas tester le bon fonctionnement de toutes les allumettes d'une production car dans ce cas tester une allumette amène à la détruire...

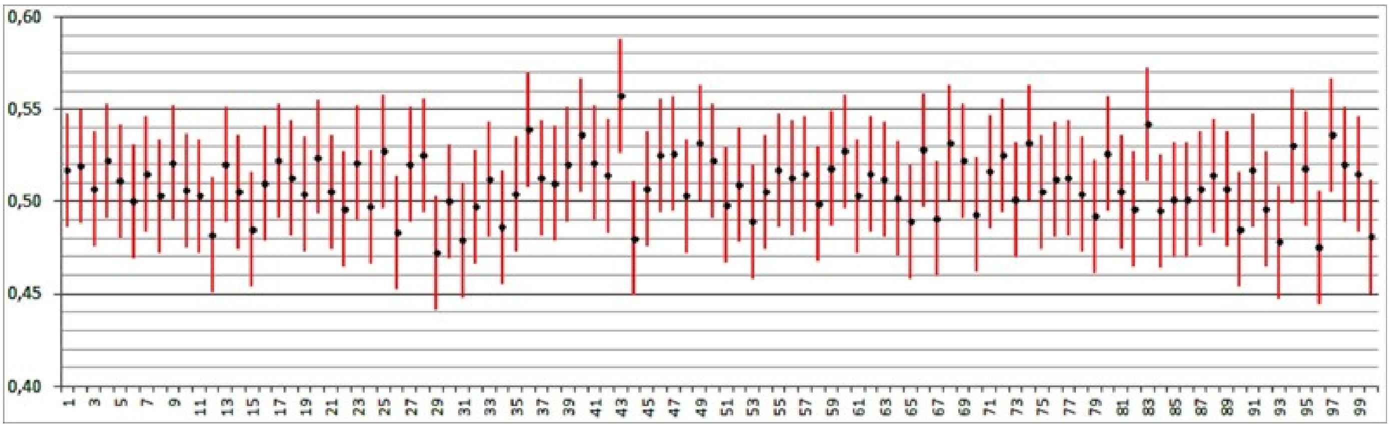
DÉFINITION.
Soit f la fréquence observée d'un caractère dans un échantillon de taille n extrait d'une population dans laquelle la proportion de ce caractère est p (inconnue).
L'intervalle $\left[f-\frac{1}{\sqrt{n}}; f+\frac{1}{\sqrt{n}}\right]$ est ***un intervalle de confiance de la proportion p au niveau de confiance de 0,95***.

On admet que l'on « peut » utiliser cet intervalle lorsque :

$$n \geqslant 30 \quad ; \quad n f \geqslant 5 \quad ; \quad n(1-f) \geqslant 5 .$$

(les conditions sur p de l'int. de fluct. sont appliquées à f)

ATTENTION : à chaque tirage d'un échantillon, on obtient un intervalle de confiance différent.
Et un intervalle de confiance peut ne pas contenir p . Un exemple :



Il s'agit d'une suite de 100 intervalles de confiance au niveau de confiance 0,95, dont chacun a été calculé sur un échantillon de taille 1000 simulé à partir de ce que fut le score de Barack Obama à l'élection présidentielle aux U.S.A. en 2012, à savoir 0,51 : dans cette simulation on observe 96 intervalles auxquels appartient la valeur 0,51 (dont 2 auxquels elle n'appartient que « de justesse ») .

Source : http://images.math.cnrs.fr/Intervalles-de-confiance-le-debat.html

A SAVOIR

1. Il est incorrect de conclure la détermination d'un intervalle de confiance par la phrase « p a une probabilité de 0,95 d'être entre $f-\frac{1}{\sqrt{n}}$ et $f+\frac{1}{\sqrt{n}}$ » car il n'y a plus d'aléatoire à ce stade : p est inconnue mais ne varie pas !
C'est pourquoi, plutôt que de parler de « probabilité », on a choisi le mot « confiance » :

« L'intervalle $\left[f-\frac{1}{\sqrt{n}}; f+\frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance de la proportion inconnue p au niveau de confiance 0,95 ».

2. La précision de l'estimation est donnée par l'amplitude de l'intervalle $\left[f-\frac{1}{\sqrt{n}}; f+\frac{1}{\sqrt{n}}\right]$, qui est $\frac{2}{\sqrt{n}}$.

Cette valeur dépend donc uniquement de n, pas de la taille de la population totale[2].

Si on veut une précision de a, on cherche à partir de quel n on a : $\frac{2}{\sqrt{n}} \leqslant a$.

Or : $\frac{2}{\sqrt{n}} \leqslant a \Leftrightarrow n \geqslant \frac{4}{a^2}$.

Voici quelques exemples :

a	0,1	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01
n					1112					

Donc, avec un niveau de confiance de 0,95, pour obtenir un intervalle de confiance d'amplitude 0,06, il faut un échantillon de taille 1112 au moins.

III. Complément : un autre intervalle de confiance

Dans les démonstrations, nous avons utilisé le résultat suivant, facile à démontrer :

$$F_n \in \left[p-\frac{1}{\sqrt{n}}; p+\frac{1}{\sqrt{n}}\right] \Leftrightarrow p \in \left[\mathrm{F}_n-\frac{1}{\sqrt{n}}; \mathrm{F}_n+\frac{1}{\sqrt{n}}\right].$$

Mais si on utilise l'intervalle de fluctuation de Terminale, c'est beaucoup plus difficile d'encadrer p …

En effet, écrire $\mathrm{F}_n \in \left[p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ ne permet pas d'encadrer p.

D'après le programme scolaire, « il est important de noter que, dans d'autres champs disciplinaires, on utilise l'intervalle $\left[f-1{,}96\frac{\sqrt{f(1-f)}}{\sqrt{n}}; f+1{,}96\frac{\sqrt{f(1-f)}}{\sqrt{n}}\right]$, qu'il n'est pas possible de justifier en Terminale ».

2 Ce qui peut étonner… Mais comme l'a dit Jean-Louis Boursin dans « les structures du hasard » : pour goûter un plat, il suffit d'en goûter une petite quantité ; cette quantité ne dépend pas de la taille du récipient (mais il faut néanmoins avoir bien mélangé) !