# les fiches cuisine de tonton lulu*

## les sondages
### cuisine "maison" ou expérience scientifique ?

Les sondages font désormais partie de notre quotidien : ils fleurissent en toutes saisons, dans toutes les branches de l'activité humaine.

Ils sont la réponse moderne à un fantasme ancien, prédire l'avenir : en effet, de tous temps, les "grands" de ce monde ont eu besoin d'une science divinatoire ; pendant des siècles, ils ont fait appel aux devins, aux astrologues et aujourd'hui, ils sollicitent les instituts de sondages.

Mais si les sondages s'inscrivent dans un cadre mathématique rigoureux — la théorie de l'échantillonnage —, leur fabrication, leur interprétation continuent-elles d'obéir à cette même rigueur ?

**Première partie :**

### LA THÉORIE : les différentes façons de goûter une soupe !

Le principe du sondage est très simple : faute de pouvoir étudier la totalité d'une population, on prélève un échantillon qu'on analyse, et on suppose que le résultat est valable pour l'ensemble de la population.

* Cette fiche ne figure pas dans le recueil de 13 "Fiches-Cuisine" paru chez Cedic-Nathan sous le titre *"Comment réussir le triangle quelconque... et douze autres friandises"*.

Prenons l'exemple d'une soupe contenant des pois chiches et des haricots ; on prélève une cuiller dans la marmite : la proportion de pois chiches est-elle la même dans la cuiller et dans la marmite ?

— "Ptêt ben qu'oui, ptêt ben qu'non", "Ça dépend."
— C'est probablement vrai ou presque".

En d'autres termes, nous voici dans la théorie des probabilités !

**A. La cuiller à café dans le bol de soupe...**

Commençons par étudier un modèle réduit : un bol de soupe dans lequel nagent 12 pois chiches et 4 haricots.

Supposons qu'avec une petite cuiller, on prélève un échantillon de deux légumes : trois "cas de figure" sont possibles :

— deux pois chiches ;
— un pois chiche et un haricot ;
— deux haricots.

On peut attribuer à chacune de ces trois éventualités, une probabilité : pour cela, on raisonne globalement (et abstraitement) : on calcule d'abord combien d'échantillons différents sont possibles, chaque pois chiche et chaque haricot étant considéré comme un individu distinct. C'est un calcul classique en mathématiques, dont je me contenterai de donner le résultat général.

Si on appelle $C_n^p$ le nombre de façons de constituer un échantillon de p individus prélevés dans une population de n individus,

$$C_n^p = \frac{n}{p} \times \frac{n-1}{p-1} \times \frac{n-2}{p-2} \times \ldots \times \frac{n-(p-1)}{1} \quad \text{ou encore} \quad \frac{n!}{p!(n-p)!}$$

Dans notre cas (16 légumes), il y a donc $C_{16}^2$ échantillons possibles, soit $\frac{16}{2} \times \frac{15}{1} = 120$.

Ensuite, on calcule combien, parmi ces 120 échantillons possibles, contiennent 2 pois chiches.

Pour constituer un échantillon de deux pois chiches, tout revient à prélever deux pois chiches parmi les douze pois chiches : on peut faire abstraction des haricots. C'est le même résultat mathématique que ci-dessus avec 12 à la place de 16 :

$$C_{12}^2 = \frac{12}{2} \times \frac{11}{1} = 66 .$$

Parmi les 120 échantillons possibles, 66 contiennent deux pois chiches. On dit alors que la probabilité de l'éventualité "2 pois chiches" vaut 66/120 = 0,55.

On raisonne de la même façon pour l'éventualité "deux haricots" : il y a $C_4^2 = \frac{4}{2} \times \frac{3}{1} = 6$ échantillons contenant 2 haricots.

La probabilité vaut $6/120 = 0,05$ .

Reste le cas "un pois chiche et un haricot" ; il y a 12 possibilités pour le pois chiche et pour chacune de ces probabilités, quatre choix possibles pour le haricot soit en tout $12 \times 4 = 48$ échantillons "mixtes".
La probabilité de cette éventualité vaut donc $48/120 = 0,4$.

On peut donc dresser le tableau suivant :

| | | | |
|---|---|---|---|
| nombre de pois chiches dans dans l'échantillon | 0 | 1 | 2 |
| avec la probabilité | 0,05 | 0,4 | 0,55 |

On remarque que la somme des probabilités vaut 1, ce qui est normal. En divisant par 2 la première ligne du tableau, on obtient :

| | | | |
|---|---|---|---|
| la proportion de pois chiches dans l'échantillon | 0 | 0,5 | 1 |
| avec la probabilité | 0,05 | 0,4 | 0,55 |

Or la proportion "véritable" (celle du bol) est 0,75. Il y a donc une probabilité 0,4 qu'on soit à $0,5 - 0,75 = -0,25$ de la proportion réelle et une probabilité 0,55 qu'on soit à $1 - 0,75 = +0,25$ de la proportion réelle.

On obtient une "fourchette" : il y a une probabilité $0,4 + 0,55 = 0,95$ pour qu'on soit à 0,25 en plus ou en moins de la proportion réelle.

Ou encore :

**Il y a 95 chances sur 100 pour que la proportion de l'échantillon soit égale à celle de la population avec une précision de $\pm 0,25$.**

Si la précision n'est pas extraordinaire, la formulation de cette conclusion est importante : c'est exactement de la même façon que s'énoncera, en toute rigueur, le résultat d'un sondage.

Pour améliorer la précision, il faut augmenter la taille de l'échantillon: reprenons nos calculs pour le bol de soupe, avec une cuiller un peu plus grande.

**B. La cuiller à soupe dans le bol... de soupe !**

Supposons à présent que notre cuiller contienne 4 légumes :

*Nombre d'échantillons possibles :*

$$\mathcal{C}_{16}^{4} = \frac{16}{4} \times \frac{15}{3} \times \frac{14}{2} \times \frac{13}{1} = 1820$$

*Echantillons à 4 pois chiches :* il y en a

$$\mathcal{C}_{12}^{4} = \frac{12}{4} \times \frac{11}{3} \times \frac{10}{2} \times \frac{9}{1} = 495$$

Soit une probabilité de $495/1820 \simeq 0,27$, pour une proportion $4/4 = 1$ de pois chiches.

*Echantillons à 3 pois chiches et 1 haricot :*

4 choix pour le haricot puis $\mathcal{C}_{12}^{3} = 220$ pour les pois chiches soit $4 \times 220 = 880$ échantillons. D'où une probabilité $880/1820 \simeq 0,48$ pour une proportion $3/4 = 0,75$ de pois chiches.

*Echantillons à 2 pois chiches et 2 haricots :*

$\mathcal{C}_{12}^{2} = 66$ choix pour les pois chiches, puis $\mathcal{C}_{4}^{2} = 6$ choix pour les haricots : soit $6 \times 66 = 396$ échantillons et une probabilité $396/1820 \simeq 0,22$ pour une proportion 0,5 de pois chiches.

*Echantillons à 1 pois chiche et 3 haricots :*

12 choix pour le pois chiche et $\mathcal{C}_{4}^{3} = 4$ pour les haricots, soit 48 échantillons et une probabilité $88/1820 \simeq 0,03$ pour une proportion 0,25.

Enfin, il y a une seule façon de choisir les quatre haricots du bol : la probabilité 1/1820 est négligeable par rapport aux autres.

Résumons ces calculs par un tableau :

| proportion de pois chiches | 0 | 0,25 | 0,50 | 0,75 | 1 |
|---|---|---|---|---|---|
| probabilité | 0 | 0,03 | 0,22 | 0,48 | 0,27 |

*Conclusion*

> Cette fois, il y a $22 + 48 + 27 = 97$ chances sur 100 pour que la proportion de l'échantillon soit égale à celle de la population, avec une précision de $\pm 0,25$.

Si l'on n'a pas gagné en précision, on a gagné en vraisemblance, le terme mathématique exact est "confiance" : 97 contre 95.

Mais en observant le tableau, on peut gagner en précision :

> **Il y a 48 chances sur 100 pour que la proportion de l'échantillon soit exactement celle de la population.**

Belle performance, "payée" par une forte perte de confiance: 48 contre 97.

Cette dialectique précision/confiance se retrouvera dans l'interprétation d'un sondage: on devra choisir un compromis satisfaisant.

### C. La soupe est-elle bonne?

Jusqu'à présent, on a supposé connue la proportion de pois chiches dans le bol (0,75), et on a étudié les proportions possibles d'un échantillon.

Mais dans un sondage, c'est tout le contraire: on connaît la proportion de l'échantillon prélevé et on cherche la proportion de la "population" totale. Supposons, par exemple, que la cuiller prélevée contienne 3 pois chiches et 1 haricot et qu'on ne connaisse pas la proportion dans le bol: est-on sûr que celle-ci vaut 0,75 ?

NON.

Il est possible que ce soit 0,75... avec une certaine probabilité. Mais il est possible que ce soit une autre valeur... avec une autre probabilité. Et on veut calculer ces probabilités.

Pour cela, il suffit de reprendre la méthode des paragraphes précédents, mais en inversant les rôles du bol et de la cuiller: on considère toutes les proportions possibles dans le bol de 16 légumes, avant prélèvement de l'échantillon. Puisque la cuiller contient 3 pois chiches et 1 haricot, il y avait au départ, dans le bol, au minimum 3 pois chiches et au maximum 15.

S'il y en avait 3 et donc 13 haricots, il y aurait 13 échantillons possibles à 3 pois chiches et 1 haricot.

S'il y en avait 4 et donc 12 haricots, il y aurait eu $\mathcal{C}_4^3$ choix pour les pois chiches puis 12 choix pour le haricot. Comme $\mathcal{C}_4^3 = 4$, il y aurait eu $4 \times 12 = 48$ échantillons possibles à 3 pois chiches et 1 haricot.

Continuons... et présentons les résultats en tableau:

| si le nombre initial de pois chiches était | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| celui des haricots était | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
| la proportion initiale était | 0,19 | 0,25 | 0,31 | 0,375 | 0,44 | 0,5 | 0,57 | 0,625 | 0,69 | 0,75 | 0,81 | 0,875 | 0,94 |
| et le nombre d'échantillons à 3 pois et 1 haricot | 13 | 48 | 110 | 200 | 315 | 448 | 588 | 720 | 825 | 880 | 858 | 728 | 455 |

Remarquons que ce calcul est mécanique et peut être confié à une machine :
Si $k$ est le nombre de pois chiches, le nombre d'échantillons est :

$$C_k^3 \times (16-k) = \frac{k(k-1)(k-2)(16-k)}{6}$$

Faisons le total du nombre d'échantillons à 3 pois chiches et 1 haricot :

$$13 + 48 + \ldots + 455 = 6188 .$$

Divisons alors par 6188 la dernière ligne pour obtenir les probabilités cherchées en supposant ces échantillons également probables. D'où :

| la proportion initiale était | 0,19 | 0,25 | 0,31 | 0,375 | 0,44 | 0,5 | 0,57 | 0,625 | 0,69 | 0,75 | 0,81 | 0,875 | 0,94 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| avec une probabilité | 0 | 0,01 | 0,02 | 0,03 | 0,05 | 0,07 | 0,10 | 0,11 | 0,13 | 0,15 | 0,14 | 0,12 | 0,07 |

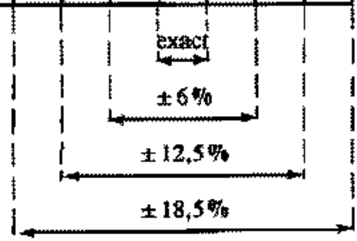
Là encore, on peut interpréter à différents niveaux de confiance et de précision :

c = 0,15

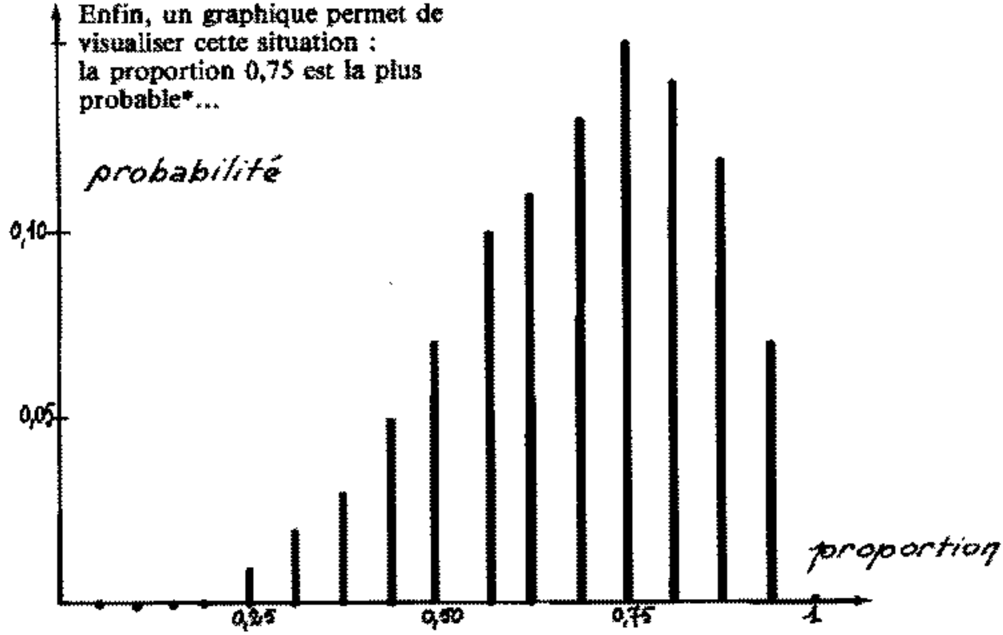
c = 0,42

c = 0,65

c = 0,82




Enfin, un graphique permet de visualiser cette situation : la proportion 0,75 est la plus probable*...




* NDLR : 0,75 est aussi la valeur de la proportion initiale qui rend maximum la probabilité $\frac{880}{1820}$ d'observer l'échantillon obtenu. On appelle cette probabilité la *vraisemblance* de l'échantillon et 0,75, l'estimation de la proportion initiale "estimateur du maximum de vraisemblance.

Ces résultats ne sont pas terribles ; pourtant, le sondage a porté sur le quart de la population... Voyons cela sur une population plus grande.

## D. La cuiller dans l'océan

Reprenons donc notre cuiller, c'est-à-dire un échantillon de taille 4, et prélevons cet échantillon dans des marmites de plus en plus grandes... Commençons comme aux paragraphes A et B par supposer la proportion dans la marmite connue et égale à 0,75; on obtient:

| | | Proportions de pois chiches dans l'échantillon | | | | |
|---|---|---|---|---|---|---|
| | Population totale | 0 | 0,25 | 0,50 | 0,75 | 1 |
| Probabilités | 16 | 0 | 0,03 | 0,22 | 0,48 | 0,27 |
| | 80 | 0 | 0,04 | 0,21 | 0,44 | 0,31 |
| | 200 | 0 | 0,04 | 0,21 | 0,43 | 0,32 |
| | 1000 | 0 | 0,05 | 0,21 | 0,42 | 0,32 |

Les probabilités sont pratiquement indépendantes de la taille de la population ! En termes mathématiques, on dit qu'il y a une convergence vers une loi de probabilité limite.

Cette loi limite est d'ailleurs connue : elle s'appelle *la loi binômiale*. Dans notre cas, elle affirme que la probabilité d'obtenir un nombre $k$ de pois chiches en prélevant un échantillon de 4 légumes vaut :

$$P(k) = C_4^k \times 0{,}75^k \times 0{,}25^{4-k}$$

où 0,75 et 0,25 sont les probabilités d'obtenir un pois chiche, ou un haricot si on prélève un seul légume. En calculant les valeurs $P(k)$ pour $k = 0, 1, 2, 3$ et 4. On peut donc ajouter une ligne à notre dernier tableau :

| | | | | | | |
|---|---|---|---|---|---|---|
| | Océan: une infinité de légumes | 0,004 | 0,0047 | 0,211 | 0,422 | 0,316 |

Donc si la taille de la population n'intervient pratiquement pas, *c'est la taille de l'échantillon qui va, seule, déterminer la qualité des résultats.*

Et si ensuite on se sert de ces résultats pour calculer, comme au paragraphe C, les probabilités d'avoir telle ou telle proportion initiale, connaissant la proportion de l'échantillon, cette propriété remarquable subsiste.

Un sondage effectué sur un échantillon de 1000 personnes donnera la même précision et la même confiance, qu'il s'agisse de Lyon ou de toute la France, de la France ou des U.S.A. !

Félicitons au passage le directeur de l'Institut XXXX pour sa mauvaise foi : à propos d'un sondage effectué à Lyon, et dont les résultats étaient contestés, il a déclaré :

"Entre un échantillon de 850 personnes pour un corps électoral de 800000 personnes dans la région lyonnaise, et un échantillon de 1000 personnes pour un corps électoral de 34 millions de Français, je vous laisse faire la comparaison..." (France-Inter, 22 janvier 1985).

**E. Une louche dans l'océan**

La taille de la population n'influence pratiquement pas le résultat du sondage si elle est grande ; ce n'est en fait pas étonnant et rejoint le bon sens. Pour goûter une soupe, on en goûte une cuiller, quelle que soit la quantité préparée. Il faut simplement que la soupe soit bien brassée !

Quant à la taille de l'échantillon, on a vu, par exemple (§ B), qu'un échantillon plus gros amenait de meilleurs résultats : là encore, ce n'est que le bon sens.

Néanmoins, il est utile de savoir à l'avance sur quelle confiance et quelle précision on peut compter, avec un échantillon de taille donnée. En d'autres termes, comment varient la confiance et la précision en fonction de la taille de l'échantillon ?

C'est une question classique en théorie de l'échantillonnage. C'est même une question d'autant plus importante qu'elle se pose chaque fois qu'on fait un sondage, depuis les intentions de vote jusqu'au contrôle de qualité en sortie d'usine (1).

La réponse à cette question passe par la *"loi des grands nombres"* qui peut s'énoncer ainsi :

> Si on prélève un échantillon de taille $n$ dans une population, alors pour chaque valeur $c$ de la confiance, il existe un nombre $h$ tel que la précision (?) soit inférieure à $\frac{h}{2\sqrt{n}}$ avec la confiance $c$.
>
> En d'autres termes, si $p$ est la proportion observée sur l'échantillon, la proportion réelle de la population est $p$ à $\pm \frac{h}{2\sqrt{n}}$ près, avec la probabilité $c$.

(1) Le contrôle d'un échantillon en bout de chaîne doit nous renseigner de façon fiable sur la proportion de produits défectueux mis en vente, sans coûter trop cher...
Un exemple est traité dans mon livre *"Maths au jour le jour"*, édition CEDIC, p. 31.

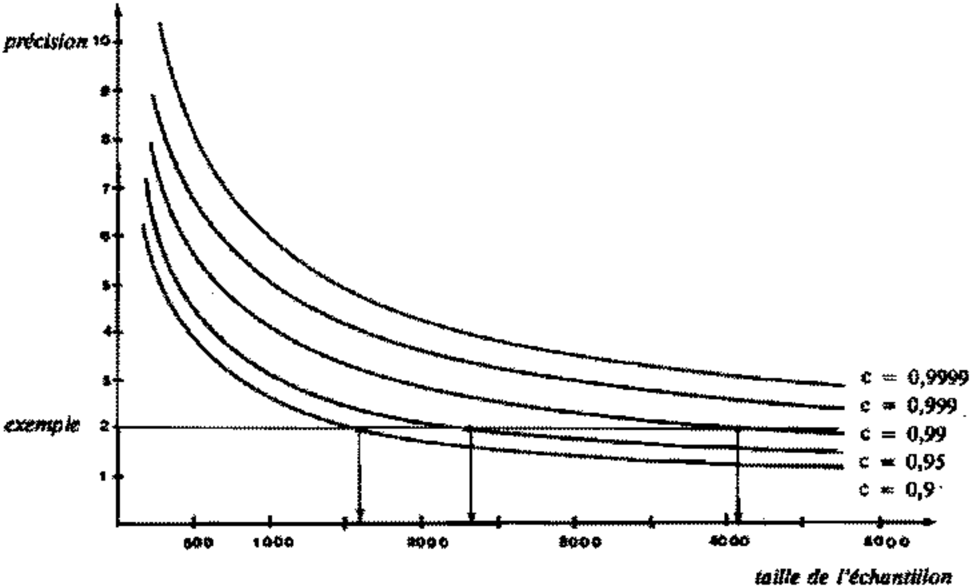
Le nombre $h$ peut se calculer à partir de $c$ par la jolie formule que voici :

$$\frac{2}{\sqrt{2\pi}} \int_h^{\infty} e^{-\frac{x^2}{2}}\, dx = 1 - c$$

Si cette formule ne vous renseigne pas (!), vous consulterez, comme moi, un tableau comme celui-ci :

| confiance $c$ | 0,9 | 0,95 | 0,99 | 0,995 | 0,999 | 0,9999 |
|---|---|---|---|---|---|---|
| valeur de $h$ | 1,645 | 1,960 | 2,576 | 2,808 | 3,291 | 3,891 |

Pour mieux comprendre les conséquences de la "loi des grands nombres", observons les courbes liant la précision à la taille de l'échantillon:




A titre d'exemple, on a cherché graphiquement quelle était la taille de l'échantillon pour être sûr d'une précision à ±2% :

— à la confiance 0,9, il suffit d'interroger 1700 personnes,
— à la confiance 0,95, il faut en interroger 2400,
— et à la confiance 0,99, il en faut 4150.

Pour les confiances plus fortes, on est trop loin... en dehors du graphique.

La pratique des instituts de sondages est de travailler avec la confiance 0,95.

Par exemple, un sondage réalisé sur un échantillon de 1000 personnes donne, à la confiance 0,95, une précision de $\pm 3,1\%$, soit une "fourchette" de plus de 6 points ! et avec 5 chances sur 100 d'être en dehors !

Dans ces conditions, à entendre les commentateurs politiques ou les chefs de parti analyser les résultats de tels sondages à un ou deux points près, on peut être *a priori* sceptique.

**Deuxième partie**

## LA PRATIQUE : la recette de la soupe humaine !

Après avoir passé en revue les différents aspects théoriques du sondage, venons en aux aspects pratiques : pour cela, suivons étape par étape la réalisation d'un sondage.

### A. Goûtez...

#### 1. *Il faut d'abord brasser*

Tous les raisonnements théoriques étaient basés sur le prélèvement au hasard d'un échantillon dans la population. Mais, dans la pratique, pour tirer au sort par exemple 1000 personnes parmi la population française, il faut d'abord posséder une liste de tous les français. Seul l'I.N.S.E.E. possède une telle liste. Ensuite, il faut effectivement aller questionner les heureux élus, où qu'ils habitent, et quelles que soient les heures où ils sont disponibles. Enfin, il faut qu'ils acceptent de répondre.

Cela fait beaucoup pour les instituts de sondages qui utilisent plutôt des échantillons dits "représentatifs" en fonction des questions qu'ils vont poser : par exemple, pour sonder des intentions de vote, on va d'abord choisir des facteurs pertinents (âge, profession, ...), puis on va constituer un échantillon contenant les mêmes proportions que la population totale, pour chaque facteur : autant de jeunes de 18 à 20 ans, autant de cadres, etc.

Cela s'appelle la *méthode des quotas* : chaque enquêteur a son quota de mères de familles, de célibataires... qu'il doit rencontrer et interroger. Ce n'est pas sans poser des problèmes pratiques : allez donc trouver un employé de bureau à son domicile... aux heures de bureau !

Pourtant, si le hasard est malmené par cette méthode, la théorie des probabilités montre qu'un échantillon représentatif donne des résultats aussi bons qu'un échantillon aléatoire, si on a effectivement tiré au sort à l'intérieur de chaque catégorie : la perte de performance est compensée par un gain d'information, apporté par les quotas.

A ce niveau, c'est donc la conscience professionnelle des enquêteurs qui va influencer la qualité du sondage. Malheureusement, leur formation et leur rémunération ne vont pas toujours dans ce sens...

## 2. *La cuiller doit être propre*

Une fois trouvé l'individu représentatif, il reste à le questionner. Mais il faut que les questions amènent des réponses exploitables, ce qui pose des problèmes de forme.

D'abord les questions doivent être claires, avoir un sens immédiat, et le même sens pour tous les gens interrogés : les mots doivent être soigneusement choisis, ne pas être ambigus ou prêter à interprétation. Par exemple, on ne demandera pas "Sortez-vous souvent ?", mais "Combient de fois allez-vous : a) au cinéma ; b) au théâtre ; c) dîner chez des amis ; ..." et la réponse indiquera le nombre de fois par mois.

Ensuite, les questions doivent être neutres, c'est-à-dire ne pas inciter à choisir une réponse plutôt qu'une autre ; efficaces, c'est-à-dire aboutir à une réponse : s'il est facile à demander à quelqu'un la puissance de sa voiture, il est beaucoup plus délicat de lui faire donner son opinion sur une question d'actualité, c'est pourquoi la forme de la question se complique :

— on passe de "Etes-vous pour ou contre ?"
— à "Etes-vous plutôt pour, indifférent, plutôt contre ?"
— puis à "Etes-vous a) tout à fait pour ; b) plutôt pour ;
c) indifférent ; d) plutôt contre ; e) tout à fait contre" ? ;
— et enfin, sommet de subtilité :
"Monsieur Untel a dit qu'il pensait que..."
Etes-vous a) tout à fait d'accord ; b) plutôt d'accord ;
c) indifférent ; d) plutôt pas d'accord ;
e) pas du tout d'accord ?".

Et on observe que, plus l'échelle des réponses possibles est étendue, plus les réponses réelles se répartissent selon une courbe en cloche, les gens se situant d'instinct au centre de l'échelle ou en tout cas pas trop loin.

De la même façon, les gens sont "plutôt pour" que "plutôt contre" : sachant cela, il suffit de poser la question dans le "bon" sens.

Il est d'ailleurs facile, si au contraire on ne veut pas poser de questions neutres, d'orienter plus ou moins visiblement les réponses : en voici des exemples sur un même thème (entre parenthèses, la réponse majoritaire) :

• "Quelle sera, d'après vous, la source d'énergie la plus importante dans 10 ans ?" (réponse : nucléaire).

• "Quelle source d'énergie souhaitez-vous voir développer d'ici dix ans ?" (réponse : solaire).

• "Etes-vous pour ou contre le développement des centrales nucléaires ?" (réponse : contre).

• "Etes-vous d'accord avec la nécessité de développer les centrales nucléaires ?" (réponse : d'accord).

En outre, on peut faire précéder les questions de préambules significatifs, par exemple, sur le même thème : "Il n'y a eu, au cours des dix dernières années, aucun accident mortel dû aux centrales nucléaires. Dans ces conditions, êtes-vous... ?" ou bien "Bien que le problème du stockage des déchets nucléaires ne soit pas réglé, êtes-vous... ?".

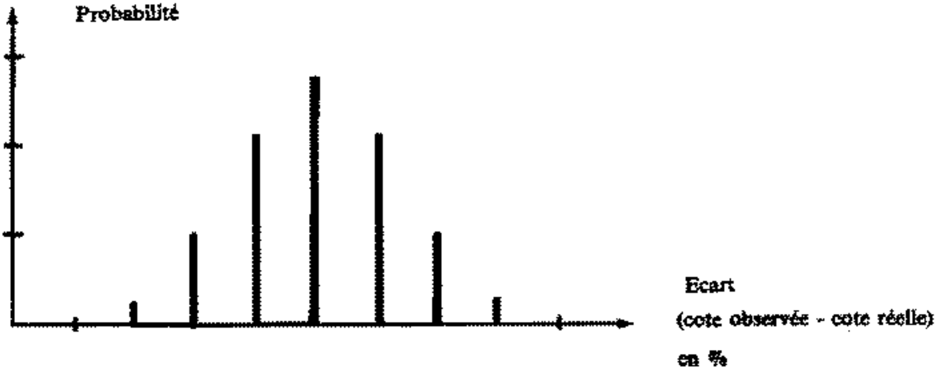
On pourrait multiplier les exemples à l'infini ; le vrai problème est celui de l'interprétation des résultats.

**B. Goûtez souvent...**

Pour vérifier la cuisson, on goûte souvent : pour étudier l'évolution de l'opinion, on sonde souvent. Et on compare les résultats : la "cote de popularité" des personnages politiques en est un exemple typique.

Mais on a vu que le résultat d'un sondage est en réalité une fourchette, assortie d'une probabilité. La question qui se pose alors est la suivante : quelle variation de la fourchette indique une variation de la proportion réelle, et avec quelle probabilité ?

Supposons, par exemple, que la "cote de popularité" de M. X soit passée de 38 % à 40 %. Or, on a vu que le résultat d'un sondage n'est pas constitué d'un seul nombre mais d'une "fourchette" et d'une "confiance" qui résultent de l'interprétation d'un tableau de probabilité (ou d'une courbe en cloche) : si par exemple, l'échantillon comprend 1000 personnes, l'écart de la "cote de popularité" réelle et la "cote de popularité" observée suit une loi de probabilité, présentée ici sous forme de graphique :




Si la cote observée est 38, on a le tableau de probabilité :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| la cote réelle est | 35 | 36 | 37 | 38 | 39 | 40 | 41 | ... |
| avec une probabilité | 0,03 | 0,10 | 0,21 | 0,28 | 0,21 | 0,10 | 0,03 | ... |

Et de la même façon, si la cote observée est 40 :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| la cote réelle est | 37 | 38 | 39 | 40 | 41 | 42 | 43 | ... |
| avec une probabilité | 0,03 | 0,10 | 0,21 | 0,28 | 0,21 | 0,10 | 0,03 | ... |

Si à présent, on se pose la question *"Quelle est la probabilité que la cote réelle ait augmenté de 2 % ?"*, il suffit d'examiner tous les cas possibles :

Notons $C_1$ et $C_2$ les cotes réelles au moment de chacun des deux sondages : $C_2 = C_1 + 2$ si :

| $C_1$ valait | $C_2$ valait | soit une probabilité de : |
|---|---|---|
| 35 | 37 | $0{,}03 \times 0{,}03 = 0{,}0009$ |
| 36 | 38 | $0{,}10 \times 0{,}10 = 0{,}0100$ |
| 37 | 39 | $0{,}21 \times 0{,}21 = 0{,}0441$ |
| 38 | 40 | $0{,}28 \times 0{,}28 = 0{,}0784$ |
| 39 | 41 | $0{,}21 \times 0{,}21 = 0{,}0441$ |
| 40 | 42 | $0{,}10 \times 0{,}10 = 0{,}0100$ |
| 41 | 43 | $0{,}03 \times 0{,}03 = 0{,}0009$ |

$$\boxed{\text{Total} = 0{,}1884 \simeq 0{,}19}$$

On peut donc conclure que la cote a augmenté de 2 points avec une probabilité de 0,19. Ce n'est pas très convaincait. Mais peut-être que la cote n'a augmenté que d'un point, ou augmenté de 3 points...

Toutes ces situations sont possibles et on peut mener des calculs analogues à celui qu'on vient de faire pour connaître les probabilités associées. Voici les résultats :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| la cote réelle a varié de | $-1$ | 0 | $+1$ | $+2$ | $+3$ | $+4$ | $+5$ | ... |
| avec la probabilité | 0,07 | 0,12 | 0,17 | 0,19 | 0,17 | 0,12 | 0,07 | ... |

on peut interpréter à différents niveaux de "confiance" (c)

c = 0,19 (de +2)
c = 0,53 (de +1 à +3)
c = 0,77 (de 0 à +4)

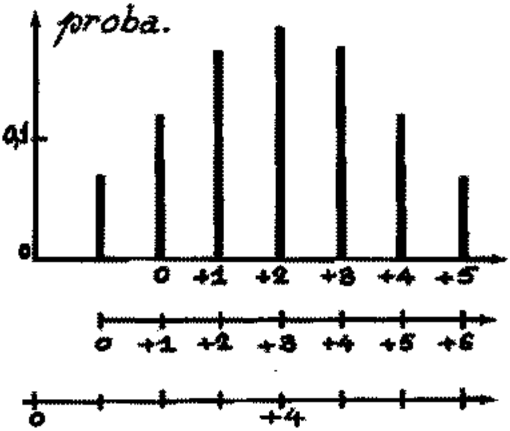
Mais il est clair que la variation observée de deux points n'est pas significative ; la variation réelle a autant de chances d'être + 2... que de n'avoir pas augmenté.

La véritable question est donc :

Quelle variation observée est-elle significative d'une variation réelle ?

En d'autres termes, à partir de quelle augmentation observée est-on sûr (au moins à la confiance 0,95) qu'il y a réellement augmentation ?

Pour cela, remarquons que le dernier tableau donne les probabilités des différents écarts possibles entre $C_1$ et $C_2$, sachant que les cotes observées diffèrent de 2%.




Si les cotes observées avaient différé de 3%, on aurait obtenu le même tableau, en décalant la première ligne d'un cran vers la gauche : 0,19 aurait été la probabilité d'une variation réelle de +3.

Cela revient à décaler, sur le graphique, le 0 de l'axe horizontal.

De combien faut-il décaler le 0 pour que celui-ci soit affecté d'une probabilité négligeable ?

Lorsque +4 est affecté de la probabilité la plus grande (0,19), la cote réelle a augmenté avec une probabilité :

$$0{,}07+0{,}12+0{,}17+0{,}19+0{,}17+0{,}12+0{,}07 = 0{,}91 .$$

Une augmentation de 4 points de la cote observée permet donc d'affirmer que la cote réelle a augmenté avec une probabilité 0,91.

Pour arriver à une probabilité 0,95, c'est-à-dire celle avec laquelle les résultats des sondages sont habituellement donnés, il faut aller encore plus loin.

***Conclusion :***

| |
|---|
| Avec un échantillon de 1000 personnes, une variation de 1, 2 ou 3 points entre deux sondages successifs n'est pas significatif. |

Ce n'est finalement pas très étonnant, puisqu'on a vu qu'avec un tel échantillon, la fourchette est de 3,1% à la confiance 0,95.

**C. Egouttez...**

La "cote de popularité" est établie à partir des réponses à une seule question : "Voulez-vous me dire si vous souhaitez voir M. X jouer un rôle important au cours des mois ou des années à venir ?" (baromètre mensuel SOFRES-Figaro Magazine). Le tri des réponses est donc simple : tant de "OUI", tant de "NON" et le reste de "non-réponses".

Mais la plupart des sondages comportent de nombreuses questions : le questionnaire contient d'ailleurs deux parties : la "signalétique" (âge, sexe, profession, etc.) et les questions propres au sondage. Le tri des réponses se fait alors en deux temps : "tri à plat" d'abord, "tri croisé" ensuite.

## 1. *La grosse passoire : le tri à plat*

On commence par compter, pour chaque question, le nombre des réponses : puis on calcule les pourcentages correspondants. Le seul problème à ce niveau est dans la prise en compte des non-réponses, cataloguées d'habitude sous le titre "sans opinion".

En effet, supposons, par exemple, qu'on ait obtenu 1/3 de "pour", 1/3 de "contre" et 1/3 de non-réponses : on pourra dire qu'un tiers de la population est "pour", mais aussi que deux tiers de la population qui ne sont pas "contre" (ainsi que deux tiers de la population qui ne sont pas "pour" !). On peut décider de ne pas tenir compte des "sans opinion" ; alors, c'est la moitié de la population qui est "pour" (et l'autre moitié qui est "contre"). Cela fait beaucoup de présentations pour un même fait !

Et si on se rappelle que la "fourchette" de précision dépend du nombre de gens qui ont fourni une réponse, il faut faire attention : dans notre exemple, notre échantillon initial est réduit d'un tiers (ceux qui n'ont pas répondu) donc la précision, qui varie comme $\frac{1}{\sqrt{n}}$ (cf 1re partie § E), est multipliée par 1,22 : si 1000 personnes avaient été interrogées, la précision passe de $\pm 3{,}1\%$ à $\pm 3{,}8\%$.

Ce phénomène apparaît aussi à l'occasion des questions "conditionnelles" auxquelles ne doivent répondre que ceux qui avaient répondu "OUI" à la question précédente :

$Q_1$ : "Etes-vous favorable à l'obligation du port de la ceinture de sécurité?"

$Q_2$ : "Si oui, pensez-vous que cette obligation doive s'appliquer :
- aux passagers avant ?
- à tous les passagers ?

Si 38% des questionnés ont répondui "oui" à $Q_1$, ils forment l'échantillon pour $Q_2$ : la précision est multipliée par 1,6 ; pour 1000 personnes au départ, elle passe de $\pm 3{,}1\%$ à $\pm 5\%$. Et plus la catégorie des gens concernés par $Q_2$ est faible, plus l'incertitude augmente...

## 2. *La passoire fine : le tri croisé*

Pour savoir qui a répondu quoi, et analyser en profondeur les résultats du sondage, on rapporche les réponses et le "signalétique" : on analyse pour chaque réponse, si un regroupement des questionnés est pertinent : les "employés" ont-ils tendance à répondre comme ceci ou comme cela ? Les "18-25 ans" se distinguent-ils dans telle réponse à telle question ?...

Pour une étude encore plus fine, on rapproche les réponses entre elles : on dispose pour cela d'un outil mathématique, d'un instrument de mesure : le "coefficient de corrélation". Il permet de détecter si on a la réponse "oui" à une question $Q_1$ en même temps que la réponse "oui" à une autre question $Q_2$ : on dit qu'il y a une corrélation entre $Q_1$ et $Q_2$ et on peut se demander pourquoi.

Par exemple, si un sondage sur le comportement des français démontre une forte corrélation entre "regarder beaucoup la télévision" et "boire beaucoup de vin", que faut-il en conclure ?

— que la télévision favorise l'alcoolisme ?
— ou que l'alcoolisme prédispose à l'abrutissement télévisuel ?

Ni l'un ni l'autre : ce ne sont pas la cause et l'effet, mais les deux effets d'une même cause — le mode de vie majoritaire en France —. Autrement dit, l'utilisation du coefficient de corrélation nécessite des précautions : si on cherche la cause d'une situation, d'une opinion, il faut d'abord examiner toutes les causes possibles, poser des questions sur toutes les causes possibles, et après seulement établir les corrélations entre les réponses à ces questions.

Les statisticiens ont développé d'autres outils pour analyser les réponses à un questionnaire, dont le principe reste de regrouper en catégories les gens et les réponses. Ces techniques utilisent l'énorme puissance de calcul des ordinateurs associée à des mathématiques sophistiquées. On arrive même aujourd'hui à analyser des réponses "ouvertes", c'est-à-dire où le questionné répond avec ses mots, ses phrases, sans être contraint de cocher des cases.

Mais la performance technique n'évacue pas les problèmes de fond...

**Troisième partie**

## LE PROBLÈME DE FOND : Y a-t-il quelque chose à goûter?

Devant l'importance prise par les sondages dans notre vie quotidienne, et au-delà des difficultés théoriques et pratiques déjà évoquées, un problème de fond est apparu : l'opinion publique existe-t-elle et, si oui, les sondages en sont-ils un reflet ? Ou, en renversant le problème : l'opinion publique existe-t-elle en dehors des sondages ?

Il faut, en effet, préciser que la très grande majorité des sondages sont des "sondages d'opinion" où les questions portent plus sur les souhaits, les intentions et les attitudes que sur les faits ou les comportements : on demandera à un individu quelle automobile il souhaite acheter et pas s'il va réellement l'acheter. L'intention n'est pas l'action.

**A. La "soupe-minute"**

Le premier aspect du problème est encore pratique : il faut répondre au questionnaire du sondage en temps limité. Pas plus l'enquêteur que l'enquêté ne veut ni ne peut y consacrer plus d'une demi-heure. Il faut donc répondre vite, avoir une opinion instantanée sur la qustion, une opinion toute prête : autant dire une opinion toute faite. Que reste-t-il de la personnalité du questionné ?

Et le choix restreint de réponses possibles ne fait qu'accentuer l'aspect caricatural de l'opinion obtenue. On ne sera pas étonné, dans ces conditions, de trouver dans un même questionnaire plusieurs réponses contradictoires. D'autant què l'enquêteur, et plus généralement l'entourage du questionné fait pression pour qu'il y ait une réponse à chaque question. L'opinion est, en quelque sorte, obligatoire : on arrive au délit de "sans opinion" !

**B. La "soupe-pression"**

La pression du milieu est plus importante qu'il n'y paraît de prime abord. L'image sociale du sondage est forte : si on pose une question, c'est qu'elle mérite d'être posée.

"Accepteriez-vous que votre fils (ou votre fille) épouse une personne de couleur ?"

Cela sous-entend que le mariage en question pose problème ; remplacez donc "de couleur" par "bretonne" ou "allemande"...

La pression impose aussi au questionné de ne pas perdre la face, de soigner son image : des "biais" systématiques sont observés quand on demande aux gens s'ils sont "extrémistes" (racisme, extrême droite... et aussi anti-communisme) ; ils sont moins nombreux à se déclarer que ceux des opinions "moyennes". Les statisticiens font alors subir aux chiffres des "redressements" calculés sur la base... de l'expérience.

Une anecdote : le CREDOC a posé, pendant plusieurs années, la question : "Avez-vous entendu parler de l'amendement Bourrier concernant la Sécurité Sociale ?" et a obtenu entre 4 et 6% de "oui". Bien entendu, cet amendement est imaginaire : encore un effet de la pression sociale...

**C. ...et les marchands de soupe !**

Enfin, dernier facteur important : l'**argent**.

Il ne faut pas oublier qu'un sondage est un produit marchand : il est vendu par un institut à un client. Ce qui amène toutes les contingences du marché :

— un sondage moins cher sera moins soigné, de moins bonne qualité ;
— un sondage plus cher souffrira de la concurrence des autres instituts ;

— les exigences du client ne sont peut-être pas l'objectivité des résultats, mais leur efficacité, surtout si le sondage est utilisé comme argument publicitaire ;
— et surtout, les résultats, l'information apportée par le sondage, sont la propriété du client. Il en dispose librement.

On est loin d'un service "public" de l'opinion publique : c'est une information partielle, colorée, qui est mise en circulation. Est-ce que, à l'instar de la presse, le pluralisme des commanditaires de sondage suffit à rétablir l'équilibre ? Quel équilibre ?

**D. Pour en finir avant l'indigestion !**

Que conclure ?

Le sondage a tous les défauts. Mais il est indispensable, ou du moins, indestructible. Alors, autant savoir
... que 1000 réponses ne sont pas 1000 opinions,
... que 1000 opinions ne se résument pas à un pourcentage,
... qu'un pourcentage ne résume pas 50 millions d'opinions.

Il n'y a pas de miracles, même en mathématiques. Ce n'est pas si grave. Tout instrument de mesure a ses limites : cela n'empêche pas de s'en servir...

**Bibliographie**

*"Les sondages d'opinion"*, H. Meynaud, D. Duclos, éd. La Découverte, 1985.

*"Les sondages, indices, statistiques, la forme scientifique du mensonge ?"* J.L. Boursin, éd. Tchou, 1978.

*"Les sondages"*, J. Gapaillard, brochure éditée par l'IREM de Nantes, 2 chemin de la Houssinière, 44072 Nantes Cedex, 1984.

*"Sondages : une forme de mensonge"*, J. Lefort in l'Ouvert n° (?).

*"Les sondages d'opinion"* in BT2 n° 91, septembre 77, édité par ICEM et CRAP.

*"Le monde aujourd'hui"*, supplément au Monde du 16-17 février 1986.

*"Théorie des sondages"*, C. Gourieroux, Economica, 1984.