

# *L'échantillon: comment ça marche*



Schweizerische Eidgenossenschaft  
Confédération suisse  
Confederazione Svizzera  
Confederaziun svizra

Département fédéral de l'intérieur DFI  
**Office fédéral de la statistique OFS**

Neuchâtel, 2009



# ***L'échantillon :*** ***comment ça marche***

*Jürg Zimmermann, Bernhard Morgenthaler, Beat Hulliger*



Un fichier Excel est disponible avec cette publication afin de vous permettre de simuler le principe de l'échantillonnage.

Les détenteurs de smartphones qui ont installé un logiciel de lecture QR sur leur téléphone peuvent accéder à ce fichier à l'aide du code QR ci-présent: Ils seront directement dirigés vers la page correspondante sur le portail statistique de l'OFS.

<b>Editeur:</b>	Office fédéral de la statistique (OFS)
<b>Complément d'information:</b>	Philippe Eichenberger, OFS, tél. 032 713 60 14 e-mail: philippe.eichenberger@bfs.admin.ch
<b>Auteurs:</b>	Jürg Zimmermann, Bernhard Morgenthaler (section Diffusion et publications, OFS), Beat Hulliger (Fachhochschule Nordwestschweiz)
<b>Diffusion:</b>	Office fédéral de la statistique, CH-2010 Neuchâtel tél. 032 713 60 60 / fax 032 713 60 61 / e-mail: order@bfs.admin.ch
<b>Numéro de commande:</b>	655-0900
<b>Prix:</b>	Gratuit
<b>Série:</b>	Statistique de la Suisse
<b>Domaine:</b>	0 Bases statistiques et produits généraux
<b>Langue du texte original:</b>	Allemand
<b>Traduction:</b>	Service linguistiques de l'OFS
<b>Graphisme/Layout:</b>	2. stock süd netthoevel & gaberthüel, Bienne
<b>Copyright:</b>	OFS, Neuchâtel 2009 La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée
<b>ISBN:</b>	3-303-00304-1

# **L'échantillon :** comment ça marche

**Quelques exemples sur le thème de la famille 4**

**Où vont-ils chercher tout ça ? 8**

**Passons à la pratique 9**

**Calculer la marge d'erreur 14**

**Qu'en est-il du hasard ? 16**

**Pas si simple ! 17**

**Tout le monde n'accepte pas de participer :**  
le cas des non-réponses **18**

**Pour conclure :** quand une enquête est-elle « représentative » ? **20**

**Annexe :** principes mathématiques à la base  
de l'échantillon **21**

**Calcul de l'écart-type et de la région de confiance 21**

**Y a-t-il une taille idéale pour l'échantillon ? 22**



## Quelques exemples sur le thème de la famille

Les quatre graphiques donnent un petit aperçu des diverses statistiques produites par l'Office fédéral de la statistique sur le thème de la famille. Ces illustrations contiennent bon nombre d'informations qui indiquent la direction prise par notre société.

GRAPHIQUE 1

### Les jeunes tardent à quitter le cocon familial ...

De nos jours, les jeunes vivent plus longtemps chez leurs parents. Ils ne deviennent pas aussi vite que leurs parents financièrement indépendants, ce qui repousse le moment où ils se mettent en ménage.

 Femmes  
 Hommes

Source: Enquête suisse sur la famille 1994/95

GRAPHIQUE 2

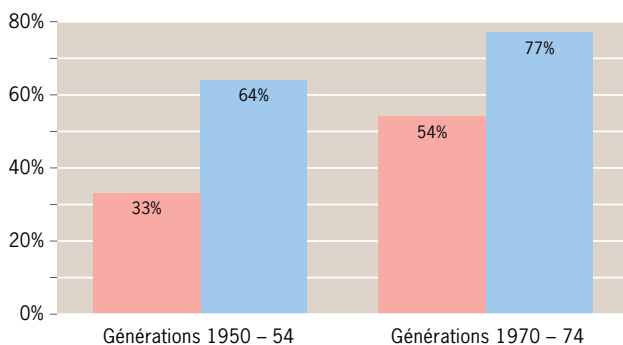
### ... mais ils n'attendent pas ensuite avant de se mettre en ménage

Une fois sortis du cocon familial, les jeunes attendent toutefois moins longtemps avant de se mettre en ménage.

Source: Enquête suisse sur la famille 1994/95

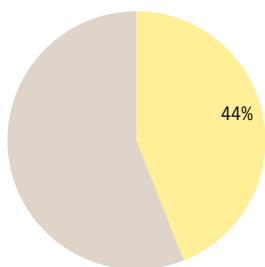
---

### Jeunes de 19 ans vivant encore chez leurs parents

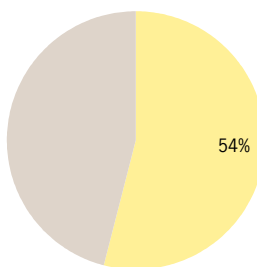


---

### Vivent avec un partenaire 3 ans après avoir quitté leurs parents



Femmes, générations 1950 – 54



Femmes, générations 1965 – 69

### Quelle est la durée d'un couple ?

Le nombre des divorces progresse fortement.



Source: Enquête suisse sur la famille 1994/95

### La famille traditionnelle, toujours la règle

La grande majorité des parents et de leurs enfants vivent aujourd'hui encore dans une famille traditionnelle (mère, père enfant/s).

Ils sont cependant nombreux à former une famille recomposée (couple vivant avec au moins un enfant issu d'une relation précédente) : c'est le cas d'un homme sur douze et d'une femme sur treize parmi les personnes de 20 à 49 ans vivant avec un enfant/des enfants :

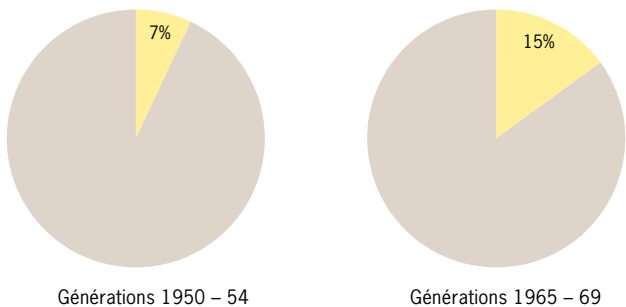
Les femmes qui vivent seules avec un enfant/des enfants sont par ailleurs loin d'être rares.

 Femmes  
 Hommes

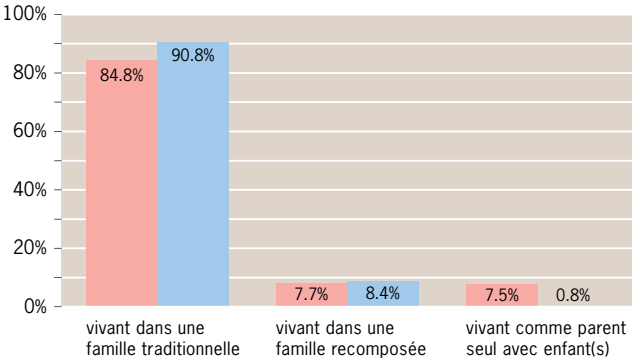
Source: Enquête suisse sur la famille 1994/95



**Jeunes de 18 ans ayant vécu le divorce de leurs parents**



**Situation familiale de la population âgée entre 20 et 49 ans**



## Où vont-ils chercher tout ça ?

« Dans les statistiques, bien sûr ». – « Mais comment font-ils ? Ils ne vont quand même pas jusqu'à interroger tous les habitants de ce pays ! ». Non, bien sûr. D'ailleurs, ce n'est généralement pas nécessaire. Et c'est là que réside l'une des clés de la statistique moderne.

La solution, c'est l'**échantillon** qui la fournit. Mais interroger des gens en les choisissant arbitrairement, n'est-ce pas le meilleur moyen d'aboutir à des erreurs ? (Les micro-trottoirs constituent un cas typique. Rien d'étonnant dans ce cas à ce que l'on entende dire que la statistique permet de prouver tout et son contraire ! )

Comment donc choisir les personnes à interroger pour que le groupe choisi (= l'échantillon) reflète fidèlement le tout (= l'ensemble de base, appelé aussi univers ou population) ?

C'est là que la statistique apporte une réponse. Cette science a montré (et démontré mathématiquement) qu'un échantillon représente très fidèlement la population lorsqu'il est tiré de manière aléatoire.

Mais avant de nous pencher sur le principe de la sélection aléatoire, voyons ce qu'il faut entendre par « obtenir des résultats représentant fidèlement la population ».

Les statisticiens parlent ici d'**estimation**. Ce faisant, ils indiquent que toutes les valeurs calculées sur la base d'un seul échantillon pour l'ensemble de la population seront entachées d'une certaine imprécision : elles présenteront un écart par rapport à la valeur réelle. C'est un fait : une estimation ne reflètera jamais exactement la réalité. Cette imprécision représente quelque chose de tout à fait concret, et sa mesure constitue l'un des champs les plus passionnants de la science statistique (nous l'aborderons au point « calculer la marge d'erreur »).

Revenons à la construction de notre échantillon. Comme nous l'avons vu, nous devons procéder de manière aléatoire.

La méthode la plus simple pour obtenir un échantillon aléatoire consiste à tirer au sort, à l'aveugle, un certain nombre d'éléments de la population, après les avoir bien mélangés, par exemple dans un sac ou une urne. Un tel échantillon porte le nom d'**échantillon aléatoire simple**. C'est à ce type d'échantillon que nous ferons allusion par la suite lorsque nous parlerons simplement d'échantillon. Dans un échantillon aléatoire simple, chaque élément de la population a exactement la même chance d'être tiré au sort. (Nous aborderons ultérieurement des cas plus difficiles, également très répandus).

## **Passons à la pratique**

Nous n'allons pas exposer les théories scientifiques et mathématiques qui fondent les méthodes statistiques. Nous aimerions plutôt illustrer le principe de l'échantillonnage, au moyen de deux exemples faciles à comprendre. Notre but est de vérifier si, en choisissant au hasard différents éléments d'un ensemble, nous obtenons une image fidèle de ce dernier. Ce faisant, nous serons forcés de constater qu'il n'est pas si facile d'utiliser la méthode du tirage au sort. Pour une population constituée de seulement 600 éléments, comme c'est le cas dans notre deuxième expérience, il nous faut déjà presque recourir à l'ordinateur. Commençons donc par une expérience que nous pouvons réaliser « au pied levé » et calculons la pointure moyenne de la classe.

**EXPÉRIENCE I : la pointure moyenne de la classe :** Imaginons que nous nous placions en rangs de six personnes dans la cour de récréation. Peu importe qu'il reste des places de libre dans la dernière rangée. Nous prenons deux dés de différentes couleurs (bleu et rouge dans notre exemple) et commençons à les lancer. Le dé bleu désigne la rangée, le rouge la colonne. Ainsi, si le dé bleu indique 2 et le rouge 3, c'est le 3<sup>e</sup> élève (à partir de la gauche) de la 2<sup>e</sup> rangée qui est tiré au sort (Attention : ne pas confondre bleu = 2 et rouge = 1 avec bleu = 1 et rouge = 2 !). L'élève en question sort du rang, les autres restent à leur place. Nous lançons les dés jusqu'à ce que nous ayons sélectionné 7 élèves (ce chiffre indique donc la taille de notre échantillon). Nous ignorons les résultats qui pointent vers une place inoccupée. Une fois que nous avons les 7 élèves, nous prenons note de leur pointure et divisons la somme par 7. La méthode d'estimation que nous appliquons consiste donc à déterminer la taille moyenne des chaussures de toute la classe

à partir de la valeur moyenne de la taille des chaussures de l'échantillon. Nous pouvons ordonner les 7 pointures de la plus grande à la plus petite et considérer pour notre estimation celle qui se trouve à mi-chemin entre les deux extrêmes, à savoir la quatrième. Cette méthode, qui nous fait recourir à la valeur médiane, présente des avantages. Ainsi, contrairement à la moyenne, la médiane ne sera pas trop affectée par une pointure sortant de l'ordinaire, comme le 53. L'inconvénient est que la médiane est généralement moins précise que la moyenne.

A présent, nous pouvons prendre note des pointures du reste de la classe et comparer la moyenne et la médiane correspondantes avec celles calculées d'après notre échantillon. A partir de notre échantillon, nous pouvons estimer d'autres informations intéressantes, comme le pourcentage des enfants uniques ou la part des élèves qui aiment le ketchup avec les frites. Et nous pouvons bien entendu répéter l'expérience, pour constater de combien la moyenne ou la médiane varie en fonction de l'échantillon.

## **EXPÉRIENCE II : *prévision des résultats d'une votation :***

Passons à présent à une expérience de plus grande envergure. Chacun peut la réaliser, à l'aide de papier quadrillé et de transparents ou à l'ordinateur. Imaginons que les 600 habitants d'une petite commune soient appelés à voter dans 4 semaines sur la rénovation de l'école. Les élèves, curieux de connaître le résultat, décident d'effectuer un sondage.

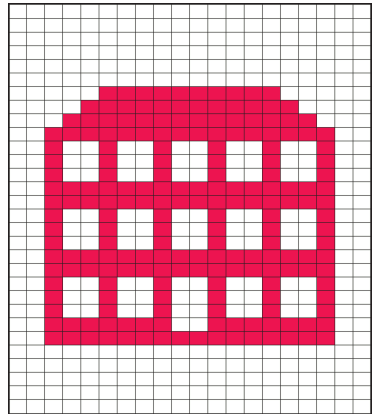
### *Comment procéder ?*

Nous avons d'abord besoin d'une population, à savoir d'un univers que nous voulons étudier. A cet effet, nous fabriquons une surface composée de 600 éléments. Une partie de ces derniers sont en rouge (dans notre sondage, cette couleur pourrait représenter toutes les voix favorables à la rénovation). Pour notre expérience, nous devons connaître le nombre véritable des champs rouges. Nous décidons d'en définir 200 comme tels (200 sur 600, soit un tiers ou 33%). Nous connaissons donc le nombre véritable de oui, que seul un échantillon nous permettra d'estimer dans la réalité. Nous tirons ensuite un échantillon aléatoire de 150 éléments. Puis nous estimons le pourcentage de oui dans la population, en comptant simplement le nombre de champs rouges dans l'échantillon, et en divisant le tout par 150.

Voici ce que cela donne:

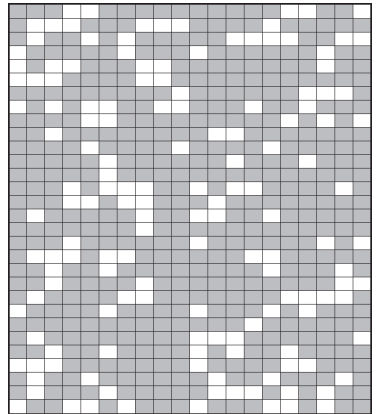
**Préparation de la population:**

Sur trois transparents, nous formons trois rectangles composés de 600 carrés (20 x 30) de taille égale. Sur le premier transparent, nous colorions 200 des 600 champs en rouge. (Nous nous sommes amusés à former une maison, mais la répartition de ces champs n'a strictement aucune importance).

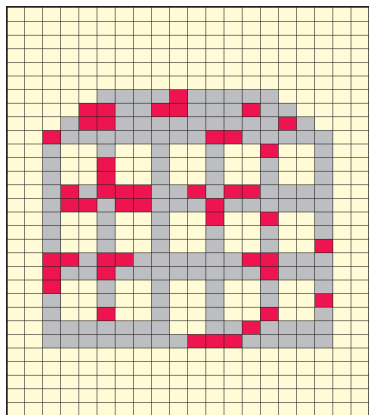


**Premier échantillon (25 % = 150 champs):**

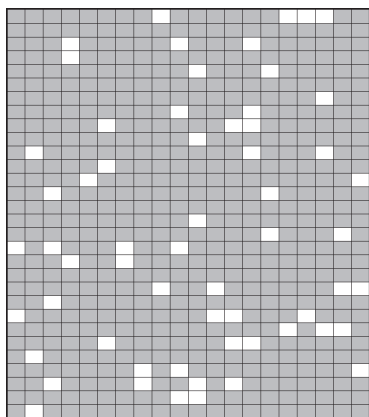
Sur le transparent suivant, nous préparons le premier échantillon : un quart seulement de la population doit en faire partie. Nous devons donc choisir 150 champs au hasard, qui resteront transparents, et recouvrir les autres (l'échantillon ne laisse transparaître que 150 éléments de la première surface). Pour cela, nous numérotons les différents champs de 1 à 600, par ligne p. ex. Maintenant, nous pourrions numéroté 600 bouts de papier, les placer dans une urne, bien la secouer, puis en tirer au hasard 150. Mais tout cela prendrait trop de temps. Nous utilisons donc le programme Excel pour créer notre échantillon aléatoire simple.



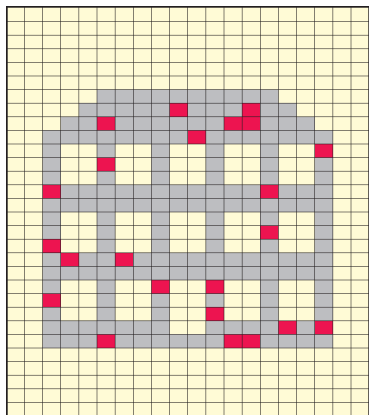
Pour les férus d'informatique : nous numérotons de 1 à 600 les cases de la première colonne d'un tableau et insérons dans celles de la deuxième colonne un nombre aléatoire situé entre 0 et 1 (fonction Excel « =ALEA() »). Nous ordonnons ensuite ces deux colonnes d'après la seconde, celle contenant les nombres aléatoires. Ce n'est pas grave si, dans Excel, les valeurs aléatoires déjà introduites changent à chaque nouvelle insertion. Nous laissons transparaître les champs contenant les 150 premiers numéros dans la première colonne : nous disposons à présent de notre échantillon.



**Première estimation :** Nous plaçons à présent le transparent contenant l'échantillon sur la population. Nous dénombrons les champs rouges visibles. Lors de notre première tentative, ils étaient au nombre de 47. Nous divisons ensuite le nombre de champs rouges par 150. Cela donne le calcul suivant :  $47/150 = 31,3 \%$ . Le résultat est donc très proche du chiffre réel (33,3 %).



**Deuxième échantillon (10 % = 60 champs) :** Nous le créons comme le premier, sauf que nous ne tirons au sort ici qu'un champ sur dix, soit 60 au total. Encore une fois, nous formons un échantillon aléatoire simple. Dans Excel, nous introduisons à nouveau des nombres aléatoires dans les champs de la deuxième colonne, ordonnons la liste selon les nombres aléatoires, puis sélectionnons les 60 premiers numéros de la première colonne. Nous laissons ces champs visibles sur le transparent.



**Deuxième estimation :** Cette fois, nous avons dénombré 23 champs. Le résultat de notre estimation est le suivant :  $23/60 = 38,3 \%$ . L'écart par rapport à la valeur réelle est un peu plus grand, mais le résultat peut encore être utilisable.

Un échantillon plus petit (p. ex. 60 champs) entraînera de fait des écarts plus grands. Chaque fois que nous répéterons l'expérience, nous obtiendrons un résultat légèrement différent, et donc un écart variant lui aussi quelque peu. En moyenne, l'écart obtenu avec les 60 champs restera plus important que celui qui résulte de l'échantillon de 150 champs.

Nous avons réalisé 60 échantillons, 30 de petite taille (60 champs), 30 de grande taille (150 champs).

**Nos résultats :**

**Nombre de champs rouges dans les 30 petits échantillons**

**(60 champs) :**

Les résultats	Résultats se situant entre		
21, 21, 17, 22, 18, 22, 17,	14 – 16	III	3
23, 17, 17, 21, 18, 25, 23,	17 – 18	IIIIIIII	10
18, 14, 25, 17, 23, 18, 20,	19 – 21	IIIIII	7
21, 16, 21, 24, 18, 27, 21,	22 – 23	IIII	6
16, 22	24 – 27	III	4

Valeur espérée : 20

Ecart moyen en termes absolus : 2,7

Ecart moyen en % : 13,5

**Nombre de champs rouges dans les 30 grands échantillons**

**(150 champs) :**

Les résultats	Résultats se situant entre		
53, 46, 46, 47, 51, 50, 50,	38 – 43	III	3
47, 44, 51, 53, 38, 64, 55,	44 – 46	IIII	5
48, 56, 46, 58, 55, 43, 59,	47 – 53	IIIIIIII	12
57, 46, 54, 54, 54, 51, 48,	54 – 56	IIII	6
47, 40	57 – 64	III	4

Valeur espérée : 50

Ecart moyen en termes absolus : 4,6

Ecart moyen en % : 9,3

Les estimations varient entre 14 et 27 pour les petits échantillons et entre 38 et 64 pour les grands échantillons. Les valeurs moyennes, de respectivement 20,1 et 50,4, sont étonnamment proches des valeurs espérées 20 et 50. Une part importante des résultats en sont encore plus rapprochés :

#### ***Ecarts observés dans les petits échantillons :***

Pour 15 estimations	l'écart est inférieur à 3 champs
pour 26 estimations	l'écart est inférieur à 5 champs
... et pour 28 des 30 estimations	l'écart reste inférieur à 6 champs

#### ***Ecarts observés dans les grands échantillons***

pour 19 estimations	l'écart est inférieur à 5 champs
pour 25 estimations	l'écart est inférieur à 8 champs
... et pour 27 des 30 estimations	l'écart reste inférieur à 10 champs

Qu'en serait-il si, au lieu de constituer un échantillon de 150 habitants dans une commune de 600 habitants, nous prenions pour référence la ville de Tokyo ? Le résultat de l'estimation ne serait-il pas inutilisable, parce que trop inexact ? Et bien non, la taille de la population n'est pas si importante : un échantillon de 150 personnes donnera des résultats presque aussi précis dans une commune de 600 habitants que dans la plus grande ville au monde ! Vous trouverez dans l'annexe davantage de détails à ce sujet.

## ***Calculer la marge d'erreur***

Notre petite expérience montre que des échantillons aléatoires simples ont de fortes chances de donner des résultats proches de la valeur réelle dans la population. Si l'on tire un grand nombre d'échantillons, la plupart donneront des résultats très proches de la valeur réelle, quelques-uns s'écarteront un peu plus de celle-ci et un petit nombre donnera des résultats plus éloignés encore. Ainsi, en tirant un seul échantillon, nous pouvons partir du principe suivant : La probabilité que le résultat de l'estimation soit très proche de la valeur réelle est très grande ; inversement, il y a peu de chance qu'il s'écarte grandement de cette valeur. Sachant cela, les statisticiens s'efforcent de déterminer de combien le résultat de l'estimation pourrait s'écarter de la valeur réelle. Pour ce faire, ils se basent sur l'**écart-type**. Nous n'allons pas



présenter ici les principes mathématiques sur lesquels ils s'appuient ; nous aimerions simplement rappeler encore une fois que l'échantillon utilisé doit être un échantillon aléatoire.

L'écart-type est donc une mesure indiquant la marge d'erreur à laquelle on peut s'attendre. Elle nous permet de décrire le degré d'imprécision de l'estimation et donc le degré d'incertitude qui subsiste. Souvent, on exprime cette incertitude par un intervalle dans lequel le résultat doit se situer avec une certaine probabilité, de 95 % p. ex. Cet intervalle est l'**intervalle de confiance**. Plus il est grand, moins l'échantillon est précis.

Dans le premier échantillon de 150 champs que nous avons constitué pour sonder la part des oui, nous avons abouti à une proportion de 31,3%. L'écart-type estimé est de 3,3 %. Nous pouvons donc dire que la valeur réelle se trouve avec une probabilité de 95 % autour de 31,3 % plus ou moins le double de l'écart-type, à savoir dans une fourchette de 24,7 % à 37,9 %. Dans le cas du petit échantillon de 60 champs, l'écart-type est plus important (5,5 %) et l'intervalle de confiance augmente en conséquence ( $38,3 \% \pm 11,0 \%$ , ou de 27,3 % à 49,3 %). (Vous trouverez dans l'annexe d'autres informations sur l'écart-type et l'intervalle de confiance).

Les statisticiens accompagnent souvent les résultats d'enquêtes par sondage des intervalles de confiance. Voici comment lire correctement le résultat de notre graphique « **La famille traditionnelle, toujours la règle** » (voir p. 6/7) :

L'intervalle de confiance à 95 % pour la proportion de femmes qui vivent seules avec des enfants parmi toutes les femmes entre 20 et 49 ans est défini par  $7,45 \% \pm 1,22 \%$ , il va donc de 6,23 % à 8,67 %.

Selon la question abordée, l'impact du résultat, l'argent à disposition, etc., le degré de précision exigé sera plus ou moins grand : des décisions impliquant de très grands montants peuvent en dépendre !

## **Qu'en est-il du hasard ?**

Une chose est sûre : les échantillons ne sont fiables que s'ils sont déterminés aléatoirement.

Mais attention : ce que nous prenons pour du hasard n'en est pas forcément.

Admettons que nous souhaitions déterminer la part des gens qui, dans notre région, sont favorables à une adhésion de la Suisse à l'UE. Faisons donc un sondage ! Nous décidons de consacrer mardi et mercredi matin à interroger les clients de trois supermarchés importants, et de décompter soigneusement les oui et les non ...

Le résultat ne peut être que faux : qui va au supermarché un mardi ou un mercredi matin ?

Il s'agit là bien entendu d'un exemple extrême. Mais il reflète les sondages effectués auprès des lecteurs de journaux populaires ou des auditeurs d'émissions de radio : seuls sont par exemple interrogés les auditeurs fidèles à une émission particulière.

Parfois, on ne se rend même pas compte qu'une partie de la population n'a aucune chance d'être prise en compte dans l'échantillon.

On en trouve un exemple dans l'histoire des Etats-Unis. Lors des élections présidentielles de 1948, les instituts de sondage prévoyaient une victoire écrasante du candidat républicain Dewey face au démocrate Truman. Pourtant, ils se sont complètement trompés. Pourquoi ? Le sondage reposait sur une enquête réalisée par téléphone.

Seulement voilà : en 1948, le téléphone était encore l'apanage des couches plutôt riches de la population, alors que Truman avait le soutien des couches les plus défavorisées.

Dans le cas de notre deuxième expérience, qui visait à déterminer le résultat de la votation, c'est comme si, dans une pièce noire, un cône de

lumière n'éclairait que le haut de notre transparent, le reste disparaissant dans l'ombre.

On ne définit donc pas un échantillon à la légère ! Il faut pour cela disposer de méthodes de tirage et de contrôle garantissant que chaque élément de la population ait une certaine probabilité (contrôlée) d'entrer dans l'échantillon.

## ***Pas si simple !***

Dans nos deux expériences, chaque élément avait la même probabilité d'entrer dans l'échantillon. En réalité, on doit souvent recourir à des échantillons constitués d'éléments qui n'ont pas les mêmes chances d'être tirés au sort. La probabilité de tirage de chaque élément doit cependant être connue de façon exacte. En d'autres termes : les spécialistes doivent souvent développer des plans d'échantillonnage complexes et des méthodes d'estimation spéciales.

Prenons comme exemple une statistique portant sur une certaine maladie de la peau en fonction des groupes d'âges, répartis par tranches de dix années : de 0 à 9 ans, de 10 à 19 ans ... jusqu'à 100 ans. Avec un échantillon aléatoire simple, il faudrait un énorme échantillon pour obtenir suffisamment de données et une estimation fiable pour les personnes de 90 à 100 ans.

Dans un tel cas, on recourra à la méthode dite de l'**échantillon stratifié**. Dans notre exemple, on constituera pour la classe d'âges des 90 à 100 ans un échantillon de même taille que pour les 60 à 69 ans, bien que les premiers soient nettement moins nombreux dans la population. La probabilité de faire partie de l'échantillon n'est dès lors plus la même dans toutes les classes d'âges (= strates). A la différence de nos deux expériences, la moyenne simple obtenue sur la base de l'échantillon ne nous fournit pas une estimation utilisable de la fréquence de la maladie de peau dans l'ensemble de la population. Nous devons pondérer les résultats, puisque les 90 à 100 ans sont surreprésentés dans l'échantillon. Cette méthode permet ainsi d'obtenir des estimations relativement précises pour des

sous-groupes particulièrement petits (les 90 à 100 ans dans notre cas), sans engendrer un coût démesuré. Mais elle suppose que l'on sache avant même de procéder au tirage à quelle classe d'âges appartient chaque élément de la population.

## ***Tout le monde n'accepte pas de participer : le cas des non-réponses***

Un autre facteur explique pourquoi les méthodes d'estimation utilisées sont souvent plus complexes que dans le cas de nos expériences: la non-réponse. En effet, il n'arrive pratiquement jamais que toutes les personnes interrogées livrent une réponse. Ainsi, lors d'une enquête sur les revenus, il y a fort à parier qu'un grand nombre des sondés refuseront de répondre. Peut-être s'agira-t-il justement des personnes qui disposent de revenus particulièrement élevés (ou particulièrement bas?). Dans un tel cas, la moyenne simple, même obtenue sur la base d'un échantillon aléatoire simple, risquerait de donner une image biaisée.

L'exemple suivant montre quelle peut être l'ampleur du phénomène de la non-réponse lors d'une enquête : il est tiré de l'indice des prix à la consommation, le « baromètre du renchérissement », qui est calculé sur la base d'un « panier-type ». Ce panier-type représente les dépenses moyennes d'un ménage suisse en denrées alimentaires, loyer, vêtements, assurances, etc. Il est établi à l'aide d'une enquête par échantillonnage auprès des ménages. Environ un tiers des ménages de l'échantillon initial ont participé à l'enquête de 2000. Deux tiers des ménages sélectionnés n'ont pas pu être atteints ou ne pouvaient pas participer (problèmes linguistiques, d'âge ou de santé), ont refusé de participer ou ont interrompu leur participation. La non-réponse est particulièrement fréquente dans certains types de ménages, notamment ceux formés de personnes âgées ou d'une jeune personne vivant seule. Ces types de ménages n'étant ainsi pas représentés correctement dans l'indice suisse, celui-ci donnerait une image biaisée du renchérissement.

Plusieurs méthodes permettent de résoudre le problème des non-réponses. On peut p. ex. comparer les estimations avec des informations sur l'ensemble de base tirées d'autres sources (registres des habitants ou autres enquêtes). Cela

doit se faire avec le plus grand soin, car il se pourrait que ces sources soient incompatibles (elles reposent sur un ensemble de base différent ou elles ont été obtenues par d'autres méthodes).

Les enquêtes dites exhaustives constituent une forme particulière de sources, dont on peut tirer des informations précieuses sur l'ensemble de base, même si elles ont été réalisées à d'autres fins. Un exemple d'enquête exhaustive est le **recensement de la population**.

Le recensement de la population nous fournit des données exactes pour notre exemple sur la consommation des ménages. Le recensement nous indique également la composition et la fréquence des différentes formes de ménage (de 1, 2 personnes ou plus, avec ou sans enfant, etc.). Ainsi, lorsque nous avons trop de réponses émanant d'un type de ménage et pas assez provenant d'un autre, nous pouvons corriger notre estimation en **pondérant** selon la taille des différents groupes, de manière à ce qu'elle corresponde à la réalité de la population.

## **Pour conclure :** *quand une enquête est-elle « représentative » ?*

Pour qu'une enquête soit jugée utilisable, elle doit satisfaire les trois conditions suivantes :

- Elle doit se fonder sur un échantillon constitué de manière **aléatoire**. De manière générale, il ne s'agira pas d'un échantillon aléatoire simple, mais d'une forme d'échantillon plus complexe. Pour que le plan d'échantillonnage soit bon, il faut avoir des connaissances suffisantes de la population.
- La **méthode d'estimation** utilisée doit permettre de tirer de l'échantillon des conclusions applicables à la population. Elle doit tenir compte de la manière dont l'échantillon a été tiré et de la non-réponse lors du relevé. Souvent, elle tient compte d'informations relatives à la population qui sont tirées de sources tierces.
- Dans la réalité, les résultats doivent être suffisamment précis pour pouvoir être utilisés. Comme nous l'avons vu, cette **précision** dépend entre autres de la taille de l'échantillon. Elle devra être plus ou moins grande, en fonction du **but de l'enquête**.

Une enquête qui satisfait à ces conditions donne une bonne image de la population. On qualifie généralement une telle enquête de représentative.

## Annexe : principes mathématiques à la base de l'échantillon

### Calcul de l'écart-type et de la région de confiance

Les statisticiens ont développé une formule permettant de calculer l'**écart-type** à partir de l'échantillon. Nous reprenons ce calcul pour notre sondage sur les résultats de la votation (chapitre « Passons à la pratique »), sans pour autant entrer dans les détails de la formule.

Données :

$N = 600$  Population

$n = 150$  Taille de l'échantillon

$m = 47$  Nombre de champs rouges dans l'échantillon

$p_m = m : n = 47 : 150 = 0,313 = 31,3\%$  Proportion de champs rouges dans l'échantillon

Voici comment est calculé l'écart-type de la proportion dans l'échantillon

$p_m$  (échantillon aléatoire simple) :

$$\hat{\sigma} = \sqrt{1 - \frac{n}{N}} \sqrt{\frac{p_m (1 - p_m)}{n - 1}} = \sqrt{1 - \frac{150}{600}} \sqrt{\frac{0,313 \times (1 - 0,313)}{149}} = 0,866 \times 0,038 = 0,033$$

Dans les cas les plus simples, la **région de confiance** se présente sous la forme d'un intervalle dit de confiance qui est défini par deux valeurs. Souvent, on calcule l'intervalle de confiance de manière à pouvoir garantir à 95 % que la valeur réelle est comprise entre ces deux valeurs (estimation le double de l'écart-type). Dans notre expérience, l'intervalle de confiance est le suivant :

$$p_m \pm 2\hat{\sigma} = 0,313 \pm 2 \times 0,033 = 31,3\% \pm 6,6\%$$

Nous pouvons de la sorte affirmer que la part réelle des champs rouges se situe avec une certitude de 95 % entre 31,3 %  $\pm$  6,6 %, à savoir entre 24,7 % et 37,9 %.

Il ne faut pas prendre cette « certitude de 95 % » au sens littéral ; de fait, soit la valeur réelle se situe à l'intérieur de l'intervalle 31,3 %  $\pm$  6,6 %, soit elle se situe à l'extérieur. Mais si nous tirions un très grand nombre d'échantillons, la

valeur réelle se situerait, pour environ 95 % des échantillons, à l'intérieur de cette marge. Plus exactement, cela signifie que l'intervalle de confiance approximatif à 95 % de la vraie proportion de champs rouges est défini par  $31,3\% \pm 6,6\%$ .

## **Y a-t-il une taille idéale pour l'échantillon ?**

Souvent, on entend dire lors de sondages qu'il existe un échantillon de taille optimale, indépendamment de la taille de la population. Ainsi, si nous effectuons un sondage auprès des populations de Suisse et des Etats-Unis, et que nous interrogeons dans chaque pays 2000 personnes, le résultat obtenu en Suisse ne serait guère plus précis que celui enregistré aux Etats-Unis, bien que la population de ce pays soit 35 fois plus importante que celle de la Suisse. Comment l'expliquer ?

Considérons la formule utilisée pour calculer l'écart-type  $\hat{\sigma}$  (voir plus haut) :  $N$ , qui représente la taille de la population, apparaît uniquement dans le premier terme :

$$\sqrt{1 - \frac{n}{N}}$$

Pour la Suisse, cette formule donne comme résultat

$$\sqrt{1 - \frac{2000}{7'000'000}} = 0,999857$$

Pour les Etats-Unis, le résultat n'est que très légèrement plus grand :

$$\sqrt{1 - \frac{2000}{250'000'000}} = 0,999996$$

Ainsi, à échantillons de taille égale, l'écart-type sera à peine plus fiable pour l'échantillon suisse que pour l'échantillon états-unien.

Les statisticiens peuvent également inverser la formule de l'écart-type pour calculer la taille d'échantillonnage garantissant la précision requise (écart-type). Pour ce faire, ils doivent toutefois savoir approximativement la part représentée en réalité par  $p_m$ .









Office fédéral de la statistique  
Espace de l'Europe 10  
2010 Neuchâtel

[www.statistique.admin.ch](http://www.statistique.admin.ch)