LES FLUCTUATIONS D'ECHANTILLONNAGE ET L'ESTIMATION D'UN POURCENTAGE PAR SONDAGE

Pour connaître la valeur d'un paramètre de santé dans une population, il n'est pas toujours possible ou nécessaire de procéder à une investigation exhaustive : on peut recourir à un *sondage,* ce qui nécessite la constitution d'un échantillon, la collecte de données et leur interprétation pour pouvoir *inférer* à la population que l'on veut étudier les résultats obtenus sur l'échantillon.

Mais l'existence d'inévitables fluctuations d'échantillonnage entraine comme corollaire la nécessité d'exprimer la valeur du paramètre de santé dans la population sous forme d'un « intervalle de confiance ».

I – FLUCTUATIONS D'ECHANTILLONNAGE D'UN POURCENTAGE OBSERVE

Soit une population de 100 000 cheptels pour laquelle le pourcentage, p, de cheptels atteints par une maladie donnée est de 20 p. cent.

❒ *Des échantillons successifs fournissent des pourcentages observés différents*

Si par tirage au sort on extrait successivement des échantillons d'un même nombre de cheptels à partir de cette population, le pourcentage de cheptels atteints dans chaque échantillon n'est pas toujours le même. Si l'on tire 1 000 échantillons de 20 cheptels à partir de cette population, le nombre de cheptels atteints dans chaque échantillon aura une distribution de fréquence observée indiquée dans le tableau Fluctuations. 1 et sur la figure Fluctuations. 1. Du fait du grand nombre d'observations, cette distribution de fréquence donne une bonne estimation de la *probabilité* de l'obtenir à nouveau lors de toute autre observation ultérieure réalisée dans les mêmes conditions.

Tableau Fluctuations. 1

Résultats obtenus lors de l'étude de 1 000 échantillons de 20 cheptels tirés dans une population où le pourcentage réel de cheptels atteints est de 20 p. cent

Nombre de cheptels atteints dans l'échantillon de 20 cheptels	Pourcentage de cheptels atteints déduit d'après le nombre de cheptels atteints trouvé dans l'échantillon	Fréquence observée sur 1 000 échantillons	Probabilité (en p. cent)
0	0	12	1,2
1	1/20 = 5	58	5,8
2	2/20 = 10	137	13,7
3	3/20 = 15	205	20,5
4	4/20 = 20	218	21,8
5	5/20 = 25	175	17,5
6	6/20 = 30	109	10,9
7	7/20 = 35	55	5,5
8	8/20 = 40	22	2,2
9	9/20 = 45	7	0,7
10	10/20 = 50	2	0,2
11	11/20 = 55	0	0,0

❒ *Les probabilités d'obtention des différents pourcentages peuvent être prédites par une loi statistique*

Les probabilités d'observation des pourcentages de cheptels atteints dans les échantillons sont calculées en acceptant l'hypothèse d'indépendance entre les cheptels tirés au sort. L'indépendance est assurée dans le cas où les tirages au sort sont faits « avec remise » de l'échantillon dans la population de façon à ne pas en modifier la composition du fait du tirage précédent (sinon, l'issue des tirages serait dépendante des tirages précédents). Dans ce cas, le nombre d'élevages atteints dans l'échantillon suit une loi binomiale de paramètres p = 20 % et n = 20, c'est-à-dire que cette loi permet de prédire les probabilités d'observer les différents résultats possibles dans le cas précis d'une population comportant 20 % d'élevages atteints et d'un échantillon de 20 individus.

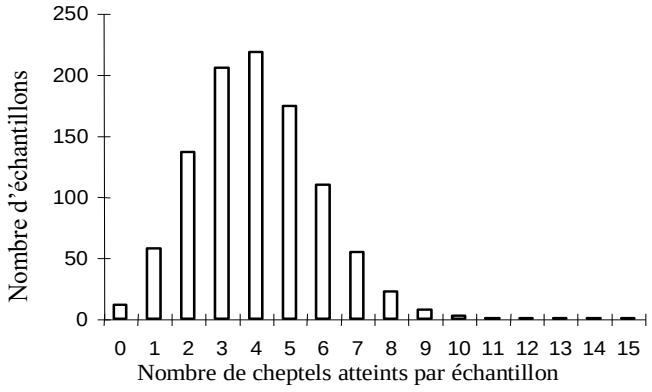
Ainsi, dans 21,8 p. cent des cas (218 échantillons sur 1 000), le nombre de cheptels atteints dans un échantillon de 20 cheptels sera de quatre, ce qui correspond exactement au pourcentage réel de cheptels atteints dans cette population (***20 p. cent***).

Mais dans un pourcentage de cas très voisin (20,5 p. cent), il sera de trois (ce qui correspondrait à un pourcentage de cheptels atteints dans la population de ***15 p. cent***, et non pas de 20 p. cent, pourcentage réel) ou de cinq (dans 17,5 p. cent des échantillons, et le pourcentage estimé est alors de ***25 p. cent*** de cheptels atteints).

Dans un pourcentage de cas d'autant plus faible que l'on s'éloigne du pourcentage réel de cheptels atteints dans la population, le nombre de cheptels trouvés atteints dans les échantillons de 20 cheptels sera faible (partie gauche de la figure Fluctuations.1) ou, au contraire, élevé (partie droite de la figure Fluctuations.1).

Figure Fluctuations. 1

Distribution de fréquence des nombres de cheptels atteints observés sur 1 000 échantillons de 20 cheptels tirés au sort dans une population où le pourcentage réel de cheptels atteints est de 20 p. cent[1]



Dans des cas extrêmes, il est possible de n'avoir aucun cheptel atteint dans un échantillon de 20 cheptels tirés au sort (dans 1,2 p. cent des échantillons) ou, au contraire, d'en avoir huit (dans 2,2 p. cent des échantillons).

[1] Noter la représentation en colonnes étroites, comme des bâtons, pour illustrer le fait que les observations correspondent à une seule valeur, sans autre valeur intermédiaire possible (voir *« Recommandations pour la réalisation et la présentation de tableaux et figures en épidémiologie animale »* sur le site de l'AEEMA, rubrique Publications).

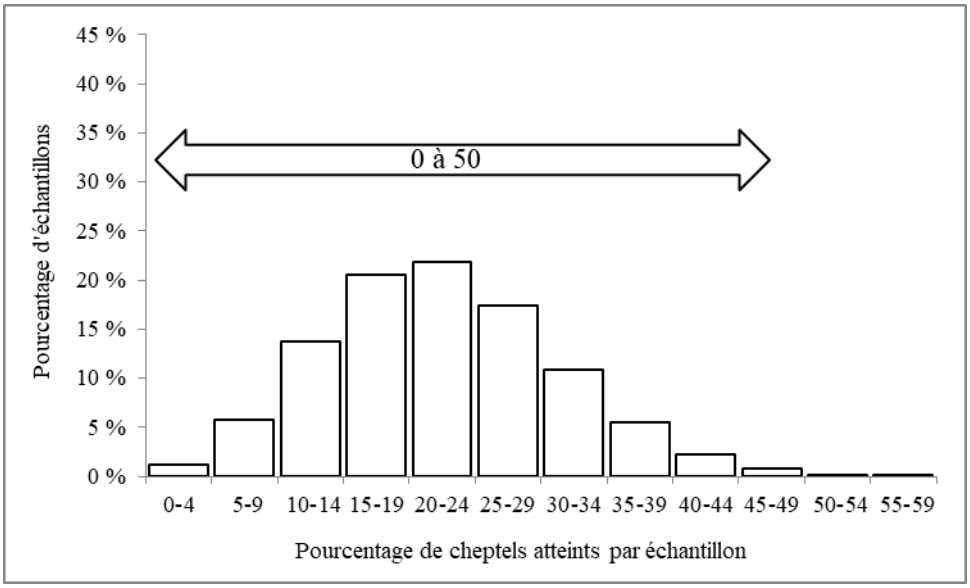
Ainsi, on constate que ***le pourcentage de cheptels atteints trouvé dans chaque échantillon est variable*** : il varie dans des proportions importantes puisque certains échantillons, certes rares, peuvent fournir un pourcentage de cheptels atteints de ***zéro p. cent*** tandis que d'autres, rares aussi, peuvent donner un pourcentage de ***50***.

Au sein de ce très large intervalle (0-50 p. cent), la distribution du nombre de cheptels atteints dans un échantillon de 20 cheptels suit une courbe « en colline » ou « en cloche », avec un sommet situé à la valeur réelle du pourcentage de cheptels atteints dans la population (*cf.* figure Fluctuations. 2).

Cette « cloche » correspond à la probabilité d'obtenir tel ou tel nombre de cheptels atteints dans un échantillon de 20 cheptels.

Figure Fluctuations. 2

Distribution de fréquence des pourcentages de cheptels atteints sur 1 000 échantillons de 20 cheptels tirés au sort dans une population où le pourcentage de cheptels atteints est de 20[2]



❒ *La valeur observée sur UN échantillon n'est pas souvent une bonne estimation de la valeur du pourcentage dans la population*

On voit donc que le pourcentage observé sur un échantillon ne correspond *exactement* au pourcentage réel que dans un faible pourcentage de cas (21,8 %) et, inversement, que dans une grande majorité des cas (78,2 p. cent = 100 p. cent – 21,8 p. cent), le pourcentage observé dans l'échantillon est différent du pourcentage réel dans la population ; *cf.* tableau Fluctuations. 1).

[2] Noter la représentation en colonnes plus larges que dans la figure Fluctuations. 1, du fait du regroupement des valeurs par incréments de 5, mais non jointives du fait de l'absence de continuité des valeurs entre deux intervalles, comme ce serait le cas pour une variable continue et avec un histogramme (voir *« Recommandations pour la réalisation et la présentation de tableaux et figures en épidémiologie animale »* sur le site de l'AEEMA, rubrique Publications).

Si donc on tirait au sort dans cette population de 100 000 cheptels **un seul** échantillon de 20 cheptels, on aurait 78,2 p. cent de risque d'arriver à une conclusion, quant au pourcentage de cheptels atteints dans la population, *inexacte* car différente du pourcentage réel.

Il n'est donc pas possible d'appliquer directement à la population la valeur trouvée sur un échantillon. La valeur trouvée se situe en fait dans une « zone » entourant la valeur réelle, avec une probabilité d'autant plus faible d'être éloignée de la valeur réelle. Ces différences, liées au ***hasard***, dans la composition de chaque échantillon, correspondent aux **fluctuations d'échantillonnage**.

❒ ***La valeur réelle est accompagnée d'une incertitude qui est fonction de la taille de l'échantillon***

Ces fluctuations sont responsables d'une **incertitude** quant à la valeur réelle du pourcentage dans la population.

Cette incertitude est comprise dans des limites qui sont fonction de la taille de l'échantillon.

Nous avons pris comme exemple, dans la population étudiée, le tirage au sort d'échantillons de 20 cheptels. On peut examiner les résultats qui seraient obtenus en tirant au sort des échantillons d'un nombre différent de cheptels, par exemple 100.

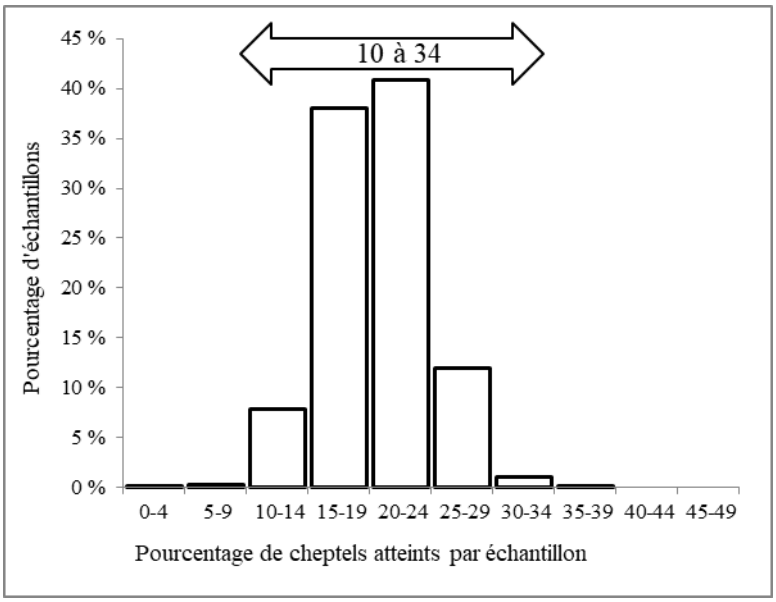
La figure Fluctuations. 3 fournit les résultats obtenus dans ce cas.

Avec des échantillons de plus grande taille (100 cheptels au lieu de 20), on observe de nouveau des fluctuations dans le pourcentage de cheptels atteints dans les échantillons étudiés. De nouveau, le pourcentage trouvé le plus souvent dans les échantillons correspond bien au pourcentage réel (classe 20-24).

La différence par rapport aux échantillons de 20 cheptels porte sur l'allure de la « colline » (*cf.* figure Fluctuations. 2) ou de la « cloche » : la colline prend l'allure ... d'une montagne, d'un pic. Sa base s'est resserrée et son sommet s'est élevé.

Figure Fluctuations. 3

Distribution de fréquence des pourcentages de cheptels atteints observés sur 1 000 échantillons de 100 cheptels tirés dans une population où le pourcentage de cheptels atteints est de 20



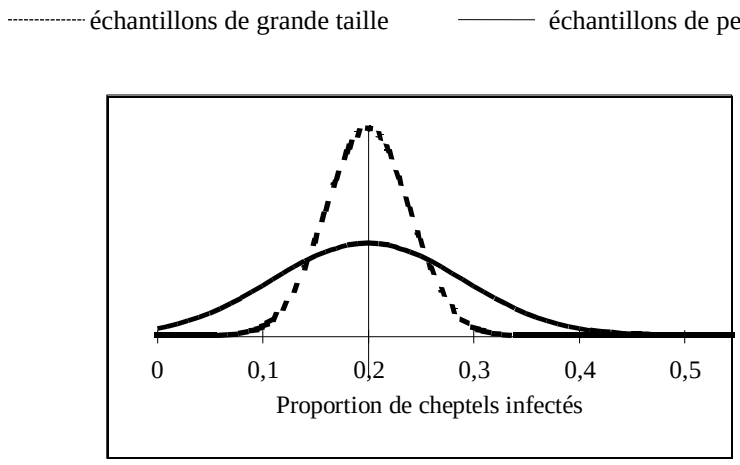
Ainsi, la zone dans laquelle se situaient les résultats obtenus avec des échantillons de 20 cheptels allait de 0 à 50 p. cent de cheptels atteints. Cette fois-ci, la distribution couvre l'étendue comprise entre 10 et 34 p. cent.

L'étude d'autres échantillons montrerait que lorsque la taille de l'échantillon augmente, l'allure de la distribution se rapproche d'une distribution dite « normale » ou gaussienne, en devenant de plus en plus étroite.

La figure Fluctuations. 4 schématise cette différence d'allure de la distribution des résultats dans deux types d'échantillons, petite taille et grande taille, par transformation des diagrammes de fréquence en courbes de distribution normale correspondantes.

Figure Fluctuations. 4

Représentation schématique de l'allure comparée des distributions observées dans des échantillons de petite et de grande taille



La fourchette de distribution des différents résultats dus aux fluctuations d'échantillonnage (ou intervalle de fluctuation) **diminue lorsque la taille de l'échantillon augmente**. Et on comprend aisément, intuitivement, que plus la taille de l'échantillon se rapproche de celle de la population, plus la valeur trouvée sera proche de la valeur réelle.

En résumé : lors de sondage, les fluctuations d'échantillonnage, dues au hasard, sont inévitables. A cause d'elles, il n'est pas possible de considérer que le résultat obtenu sur UN échantillon est le même que celui de la population d'origine de l'échantillon.

II - ESTIMATION PAR INTERVALLE D'UN POURCENTAGE

Sur un échantillon unique utilisé, le pourcentage obtenu est p_o (pour « observé »). Plutôt que de se contenter de cette valeur p_o comme estimation *ponctuelle* de la véritable valeur cherchée p, dont on a vu qu'elle avait une probabilité élevée d'être *inexacte*, on préfère calculer un intervalle dans lequel on fait le pari que la vraie valeur p a de grandes chances de se trouver. Cet intervalle calculé est appelé **intervalle de confiance** (ou fourchette) car on peut définir le degré de confiance pour que le véritable pourcentage p y soit situé.

La formule approchée définissant l'intervalle de confiance pour un risque d'erreur α (*cf.* Annexe *« Les erreurs de jugement »*) est la suivante :

$$\text{intervalle de confiance : } p_o \pm \varepsilon_\alpha \sqrt{\frac{p_o\, q_o}{n}}$$

avec p_o : pourcentage observé sur l'échantillon
q_o : complément à 1 de p_o
ε_α : écart-réduit correspondant au risque d'erreur α
n : nombre d'unités dans l'échantillon.

Les **conditions d'application** de cette formule sont :

$$\mathbf{np_0 \geq 5 \text{ et } nq_0 \geq 5}$$

☐ ***L'approximation obtenue en calculant l'intervalle de confiance*** dans lequel on espère que se trouve la véritable valeur p, à partir de p_o (qui varie d'un échantillon à un autre), ***est acceptable***. En effet, si l'échantillon est suffisamment grand (*cf.* les conditions d'application), les fluctuations d'échantillonnage restent modérées et le pourcentage inconnu p ne doit pas être très éloigné de la valeur observée p_o. D'autre part, le produit pq est assez stable car, lorsque p augmente, q diminue (et réciproquement). Pour ces deux raisons, le produit de valeur inconnue pq doit être assez voisin de p_oq_o.

Le calcul de l'intervalle de confiance d'un pourcentage observé implique tout d'abord de **fixer le seuil de risque d'erreur** α que l'on accepte de voir la véritable valeur se situer en dehors de l'intervalle de confiance.

Très souvent, ce risque est fixé à 5 p. cent. Il s'agit alors d'un intervalle de confiance à 95 p. cent, ce qui signifie que la véritable valeur p a 95 p. cent de chance de se situer dans l'intervalle de confiance ainsi calculé et 5 p. cent de risque d'être en dehors de cet intervalle.

La valeur de l'écart-réduit correspondant à un risque d'erreur α de 5 p. cent est de 1,96. En pratique, cette valeur est arrondie à 2 dans les calculs et l'intervalle de confiance à 95 p. cent comprend donc de part et d'autre de la valeur observée p_o **deux écarts-type** (*cf.* figure Fluctuations. 5).

***Exemple** : si sur un échantillon de 100 unités on obtient 23 réponses positives,*

On a : n : 100 p : 0,23 q : 0,77

$np_0 = 23$ et $nq_0 = 77$: Les conditions d'application sont donc satisfaites.

L'intervalle de confiance à 95 p. cent est : $0{,}23 \pm 2\sqrt{\frac{0{,}23 \times 0{,}77}{100}} = 0{,}23 \pm 0{,}08$

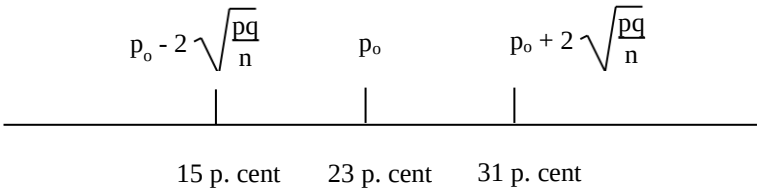
I.C. à 95 p. cent : 15 p. cent à 31 p. cent

On a 95 p. cent de chances pour que la véritable valeur de p soit située dans la fourchette 15-31 p. cent.

On peut également calculer l'intervalle de confiance à 99 p. cent. Dans ce cas, la valeur de l'écart-réduit pour un risque α de 1 p. cent est 2,57, que l'on peut arrondir à 2,6.

Figure Fluctuations. 5

Intervalle de confiance à 95 p. cent d'un pourcentage



Pour l'exemple précédent, l'intervalle de confiance à 99 p. cent est :

$$0{,}23 \pm 2{,}6\sqrt{\frac{0{,}23 \times 0{,}77}{100}} = 0{,}23 \pm 0{,}11$$

I.C. à 99 p. cent : 12 p. cent à 34 p. cent

On a 99 p. cent de chances pour que la véritable valeur de p soit située dans la fourchette 12-34 p. cent.

Bien sûr***, l'intervalle de confiance à 99 p. cent est toujours plus large que l'intervalle de confiance à 95 p. cent.***

☐ ***La formule approchée donnant l'intervalle de confiance*** pour un risque d'erreur α n'est une approximation correcte que si la ***taille de l'échantillon est suffisante***.

Pour des tailles d'échantillons plus faibles ($np_0 < 5$ et $nq_0 < 5$), on peut utiliser un calcul plus précis qui fournit un intervalle de confiance non symétrique (sauf pour un pourcentage de 50 p. cent). La formule à appliquer n'est pas simple. Le fichier Excel de l'annexe *Taille d'un échantillon aléatoire simple* permet de faire le calcul.

N.B. Lorsque la fraction de sondage $\left(\frac{n}{N}\right)$ est supérieure à 10 p. cent, l'intervalle de confiance devient :

$$p_o \pm \varepsilon_\alpha \sqrt{\left(1-\frac{n}{N}\right)\frac{p_o q_o}{n}}$$

Avec p_o : pourcentage observé sur l'échantillon
ε_α : écart-réduit correspondant au risque d'erreur α
n : nombre d'unités dans l'échantillon
N : nombre d'unités dans la population
q_o : complément à 1 de p_o

En résumé : l'utilisation des lois de probabilités appliquées à un échantillon ayant fourni un pourcentage observé permet de calculer l'intervalle de confiance dans lequel le pourcentage réel de la population a de grandes chances de se trouver. Pour calculer cet intervalle, on se sert des données caractérisant l'échantillon : la valeur du pourcentage observé (p_0) et sa taille (n).