LE CONDITIONNEMENT INVERSE :
Plus de vigilance dans l'analyse des proportions
****

Elysée Aristide
HOUNDETOUNGAN
Ingénieur Statisticien Economiste (ISE)
en formation à l'ENSAE de Dakar
email : ariel92and@gmail.com
3 avril 2015

Les données statistiques n'ont jamais autant de sens si elles ne sont pas interprétées. C'est là aussi, le rôle d'un analyste de données. Ce dernier doit toujours accompagner ses informations chiffrées d'un commentaire, faisant ainsi "parler" les chiffres.

1 Le conditionnement inverse

En statistique descriptive, une des disciplines essentielles de la statistique, la comparaison des proportions entre différentes populations peut être facilement biaisée. En effet, plusieurs analyses, par manque de vigilance, sont inclinées par ce qu'on appelle « conditionnement inverse ». Du coup, les données peuvent interpréter l'inverse de ce que l'on pourrait penser.

Considérons l'information suivante :

♣ *A l'issue d'un concours auquel 6000 candidats ont participé, 500 personnes sont retenues dont 350 hommes . . .*

La question est la suivante :

☞ Les lauréats sont composés de 350 hommes et 150 femmes. Lorsqu'on calcule les pourcentages, on obtient :

$$\mathbf{Pourc_{femme}} = \frac{\mathbf{150 \times 100}}{\mathbf{500}} = \mathbf{30} \qquad \mathbf{Pourc_{homme}} = \frac{\mathbf{350 \times 100}}{\mathbf{500}} = \mathbf{70}$$

Qu'allons-nous conclure ? les hommes ont-ils plus réussi au concours ?

C'est justement le manque de vigilance dont je parlais. Le bon statisticien n'a pas encore toutes les informations pour conclure. En effet, l'on ne devrait pas calculer la distribution du sexe parmi les lauréats. C'est ce qu'est « le conditionnement inverse ». On devrait déterminer, au sein de chaque genre (féminin, masculin), la proportion de lauréats. Peut-être que l'effectif élevé des hommes parmi les lauréats est provoqué par un effectif élevé d'hommes ayant participé au concours. Peut-être que, s'il y avait le même nombre de femmes et d'hommes au concours, on aurait un résultat contradictoire.

Supposons que l'on dispose de l'information supplémentaire ci-après :

♣ . . . *Il y avait 250 femmes au concours et donc 5750 hommes.*

Partant de cette information, on conclut que, sur 250 femmes, 150 sont admises et sur 5750 hommes 350 sont admis. C'est cette analyse qui est plus pertinente et s'interprète en termes de comparaison du taux d'admission entre femmes et hommes. Les pourcentages calculés sont :

$$\mathbf{Pourc_{femme}} = \frac{\mathbf{150 \times 100}}{\mathbf{250}} = \mathbf{60} \qquad \mathbf{Pourc_{homme}} = \frac{\mathbf{350 \times 100}}{\mathbf{5750}} = \mathbf{6.1}$$

Ainsi, 60% des femmes sont admises tandis-que le taux de réussite est environ de 6% chez les hommes : les femmes ont plus réussi au concours.

La comparaison de proportion entre deux populations devra être faite avec plus d'attention. Il faut calculer dans chaque population un taux pour que l'analyse ait un sens. Il ne faut pas

calculer la distribution du profil de la population (dans notre cas le sexe) au sein de ceux qui sont touchés par un phénomène.

Dans une analyse des taux de pauvreté, il ne faut pas calculer le pourcentage d'hommes et de femmes parmi les pauvres. Il faut plutôt déterminer le pourcentage de "pauvre", parmi les hommes d'une part, et les femmes d'autre part, pour pouvoir connaitre lequel des deux genres est plus touché par la pauvreté.

2 Insuffisance de taille

Au-delà du conditionnement inverse, la taille des populations qu'on compare peut fausser aussi les analyses. Considérons les informations suivantes :

- ♣ 1- *Malgré les résultats catastrophiques de l'examen, l'école « A » a donné un taux de réussite de 100% et l'école « B » 67% ;*
- ♣ 2- *67% de ceux qui ont contracté le virus « V1 » ont survécu et aussi 67% de ceux qui ont contracté le « V2 » ont survécu ;*

Ici le conditionnement n'est pas inversé. Il s'agit bien des proportions de personnes ayant subi un phénomène dans chaque population, qui pourraient permettre de comparer les populations. Mais nous ne pouvons pas conclure. En effet, sans une information sur les fréquences absolues (effectifs) de chaque population, les fréquences relatives (pourcentages) ne sont pas très utiles.

Dans le premier exemple, même en supposant que le résultat d'un élève à l'examen est uniquement déterminé par la performance de son école, nous ne pouvons pas conclure que l'école « A » est meilleure. **Imaginez-vous que l'école « A » a présenté un seul élève à l'examen et l'école « B » 3.** Donc le seul élève de « A » est admis tandis que 2 sur 3 sont admis dans l'école « B », ce qui donne les taux 100% et 67%. Ces effectifs sont trop faibles pour faire une comparaison. Peut-être que s'il y avait 50 élèves dans chaque école, les taux seraient de moins de 20% pour « A » et 70% pour « B ». On ne peut pas conclure aussi que les écoles sont de référence parce qu'elles donnent de bons taux de réussite malgré les résultats catastrophiques dans l'ensemble. Le même raisonnement peut être mené dans le second cas. Les informations ne sont pas suffisantes pour conclure que les deux maladies ont la même létalité. Peut-être que le pourcentage pour le virus « V1 » est calculé sur une population de grande taille et celui du virus « V2 » sur une population de petite taille, ou les deux taux sur des populations de petites tailles. On aurait une information différente en présence de grands échantillons. Il faut donc comparer les proportions entre deux populations lorsque les tailles des populations sont grandes (généralement supérieures ou égales à 25).

Il faut donc retenir que, l'analyse des données statistiques requiert plus d'attention et de vigilance. Les chiffres apportent beaucoup d'informations et peuvent aussi provoquer la prise de décisions incorrectes, lorsqu'on est un peu distrait. Pensez toujours, dans vos analyses, au conditionnement inverse et à la taille de l'échantillon !