Direction générale
Inspection générale

Note

Dossier Suivi par :
RIVIÈRE Pascal
Tél : 01 87 69 50 66
Mèl : pascal.riviere@insee.fr

Montrouge, le 17 janvier 2022
N° 2022_4/DG75-B001

Objet : Statistique et data, éléments de différenciation

Nous sommes dans un univers dans lequel les données jouent désormais un rôle croissant, et ce dans tous les domaines d'activité, au quotidien. On parle ainsi de données, de *data*, … et, parallèlement, de « chiffres », de statistiques. Insidieusement une confusion s'installe, ce qui pose problème car ce sont des notions bien distinctes.

Qu'en est-il de la différence entre donnée et statistique ?

Pour mieux l'appréhender, prenons quelques exemples de données. On peut illustrer cette notion à travers différents cas tirés de situations et domaines d'activités très variés : le montant d'une facture, le code d'activité économique d'une entreprise, la température relevée par un capteur, la consommation électrique d'un ménage, le code d'immatriculation d'un véhicule, la date de début d'un contrat, et pour finir le taux de chômage.

Pratiquement, une donnée se caractérise par trois composantes :

- un concept : le nom de la donnée, en quelque sorte,
- une valeur : par exemple 23,5°C, 45 euros, …,
- un domaine de valeurs possibles : par exemple l'ensemble des valeurs relevant de la nomenclature d'activité économique.

Une donnée « naît » dans un certain environnement, et peut être obtenue selon différentes modalités : saisie manuellement par un opérateur, déterminée de façon automatique par un capteur, calculée automatiquement à partir d'autres données. Elle peut aussi résulter d'un processus de production très complexe, avec de nombreuses étapes de traitements, manuels ou automatisés : c'est justement le cas des statistiques.

Dans les faits, l'immense majorité des données existantes dans le monde numérique sont des données individuelles, élémentaires : données relatives à une personne, une voiture, une entreprise, un événement, etc. Ainsi, elles sont le plus souvent associées à des objets à un niveau très fin (telle facture), localisés dans l'espace (tel capteur, positionné à tel endroit) et le temps (la consommation d'électricité pendant telle période). Dans le tout dernier exemple cité plus haut, le « chiffre du chômage », il s'agit au contraire d'une donnée beaucoup plus élaborée, nécessitant de s'appuyer sur d'immenses quantités de données individuelles et des calculs intermédiaires dérivés de celles-ci.

Ainsi, une statistique est-elle bien une donnée, mais une donnée très particulière :

- par sa nature : il s'agit d'une donnée agrégée et non d'une donnée élémentaire,
- par son mode d'obtention : ni saisie, ni capteur, ni calcul automatique, mais succession d'opérations de transformation.

Ces deux caractéristiques n'épuisent pas les distinctions entre les statistiques (au sens de statistiques publiques) et les données.

Lorsqu'il s'agit de statistiques *publiques*, la différence majeure réside dans leur finalité, qui est exclusivement une **finalité d'information générale**, alors que les autres données existent à des fins opérationnelles (gestion, aide à la décision, information individuelle, ...). Cette caractéristique fondamentale n'est pas sans conséquence, puisqu'il en découle des besoins de qualité et de transparence.

Pour cela, le processus de production des statistiques publiques doit obéir à un certain nombre d'exigences, indispensables pour avoir le statut de référence dans le débat public.

Plus spécifiquement, et contrairement aux données en général, la façon même de construire une statistique publique n'est pas libre, car soumise à de très nombreux contrôles :

- sur l'utilité publique de celle-ci via le Conseil National de l'Information Statistique,
- sur la qualité de la démarche à travers le Comité du label de la statistique publique : questionnaire, échantillonnage, mais aussi méthodes de traitement des données administratives et analyse de leur qualité,
- sur le respect de la confidentialité via le RGPD et le rôle du comité du secret,
- sur la conformité aux règles européennes : usage de nomenclatures standardisées, référence à des concepts reconnus (règlements européens), respect du code des bonnes pratiques (rôle de l'Autorité de la Statistique Publique).

Elle s'appuie également sur des méthodes éprouvées, partagées au niveau international : techniques d'échantillonnage, de questionnement, de contrôle de cohérence, de traitement des non-réponses ou non-déclarations, ou de calcul de précision.

Les besoins de transparence liés aux statistiques publiques engendrent d'autres exigences à satisfaire, cette fois sur les modalités de diffusion.

La mise à disposition de ces statistiques est en effet particulièrement contrainte et surveillée :

- information en amont sur les enquêtes lancées (via le CNIS, à nouveau),
- mise en ligne à la fois des statistiques elles-mêmes, mais aussi des définitions et des nomenclatures utilisées,
- règles d'embargo liées aux dates de publication,
- enfin, de façon moins formelle, on constate une demande croissante de transparence sur les algorithmes et les méthodes.

On le voit, les statistiques publiques sont donc bien des données, mais des données très particulières, complexes à la fois à travers leur mode de construction et l'environnement riche qui encadre leur élaboration et leur publicité. Or, pour la majorité des données disponibles dans l'univers numérique, qu'elles soient *big*, ou, mieux, *open*, un tel encadrement n'existe pas. Ou s'il existe, le niveau d'exigence attendu est sans commune mesure.

Le Chef de l'Inspection générale

Signé : Pascal RIVIÈRE