

No 11-522-XIF au catalogue

**La série des symposiums internationaux
de Statistique Canada - Recueil**

**Symposium 2005 : Défis
méthodologiques reliés aux
besoins futurs d'information**



2005



Statistique
Canada

Statistics
Canada

Canada

LA COMMUNICATION DE LA VARIANCE ET DES ERREURS D'ÉCHANTILLONNAGE AUX UTILISATEURS

Ed Swires-Hennessy¹

RÉSUMÉ

Les statisticiens définissent de nouveaux concepts pour communiquer les erreurs liées aux estimations. Bon nombre de ces concepts sont faciles à comprendre pour les statisticiens, mais sont bien plus difficiles à expliquer aux utilisateurs que le traditionnel intervalle de confiance. La solution proposée, pour communiquer avec des non-statisticiens, consiste à améliorer les estimations de façon à réduire au minimum la nécessité d'expliquer les erreurs. On évite ainsi de déconcerter l'utilisateur en lui fournissant trop de chiffres à comprendre.

MOTS CLÉS : concept d'erreur statistique; communication; utilisateurs.

1. EXPOSÉ

La science statistique, voire le rôle des statisticiens, consiste à établir des estimations à partir de données incomplètes. Cependant, ces estimations comportent des erreurs. L'exposé de John Wood a résumé les sources d'erreur et proposé quelques idées pour les présenter. Toutefois, ce genre de publication pose un défi aux producteurs de données.

Deux des trois exposés de la présente séance ont montré différentes façons d'amener l'utilisateur à comprendre l'erreur liée aux données; le troisième a présenté un moyen de communiquer des variances pour l'interprétation des variations et des points d'inflexion dans les enquêtes répétées. L'exposé de John Wood, présenté par Markus Šova, a fait état des diverses sources d'erreur, mais concluait que l'analyste devait envisager un compromis entre la réduction de la variance des estimations et le coût de réalisation de la réduction. Avi Singh, co-auteur d'un exposé avec Matt Westlake et Moshe Feder, a présenté un nouvel aspect du coefficient de variation ayant pour objet de résoudre les problèmes liés aux estimations autour de zéro et a proposé une nouvelle mesure : le coefficient de variation discriminant.

La majorité des données statistiques sont présentées sous forme de tableaux. Or, « extraire de l'information d'un tableau, c'est comme extraire la lumière solaire d'un concombre », ont affirmé ironiquement Farquhar et Farquhar (1891). L'extraction d'information est réputée quasi impossible, mais pourquoi? Au fond, 80 % du grand public n'entendent rien aux chiffres; en leur présentant un tableau de données, on ne leur fournit donc pas une information assimilable. Chez les journalistes (Worcester, 2004), la situation est encore pire : 95 % d'entre eux n'y entendent goutte, puisqu'ils ont étudié les arts dans une plus forte proportion que le grand public. Dans le cadre de son exposé, Avi Singh a mentionné que le coefficient de variation était « facile à comprendre ». J'aimerais bien être présent dans le bureau du premier ministre lorsqu'Avi tente de lui expliquer ce concept! Le coefficient de variation est peut-être compréhensible pour le statisticien, mais est-ce bien l'utilisateur auquel nous nous adressons?

Avant de répondre à cette question, permettez-moi d'aborder les aspects des renseignements supplémentaires que les statisticiens souhaitent diffuser avec leurs estimations. Je regroupe ici les erreurs d'échantillonnage (et autres), les estimations de la variance, les intervalles de confiance, les coefficients de variation et la nouvelle statistique proposée lors du présent symposium : le coefficient de variation discriminant. Le principal objectif de ces renseignements supplémentaires consiste à préciser l'exactitude, car en bons statisticiens, nous voulons indiquer dans quelle mesure nos estimations sont bonnes (ou mauvaises!) L'utilisateur averti peut aussi tenir compte de ces

¹ Edward Swires-Hennessy, Data Unit Wales, Columbus Walk, Cardiff, Wales, United Kingdom, CF104BY (Ed.Swires-Hennessy@dataunitwales.gov.uk).

renseignements supplémentaires pour évaluer la sensibilité d'une décision à prendre et l'incidence probable de cette décision si l'estimation se situe à la limite supérieure ou inférieure de l'intervalle de confiance. Néanmoins, lorsqu'on informe un élu chargé de prendre une décision importante en matière de dépenses qu'à l'estimation produite par les statisticiens correspond un intervalle de possibilité compris entre 30 et 36, il comprend spontanément « 33 ».

Le diffuseur de ce renseignement supplémentaire fait face à des problèmes importants. L'utilisateur n'en comprend pas l'objet ni, très souvent, les concepts. En ajoutant des chiffres à l'estimation, on ne fait que déconcerter l'utilisateur, car peu d'entre eux ont appris les concepts qui sous-tendent les renseignements supplémentaires. En fournissant une estimation accompagnée d'un intervalle, on risque d'embrouiller la prise de décision; en fournissant plus d'une estimation, chacune accompagnée d'un intervalle, on déboussole complètement l'utilisateur moyen. En outre, les statisticiens ne sont pas habitués à commenter par écrit leurs statistiques ni à décrire la fiabilité de chaque estimation.

Certains problèmes liés à la diffusion d'information sur les erreurs de mesure tiennent au fait que les concepts sont mal compris par les non-mathématiciens. Bon nombre d'estimations sont présentées sans être arrondies et les statisticiens, agissant comme des comptables, présumant qu'une plus grande précision (vraie ou fausse) s'impose; or, cela ne fait qu'ajouter à la confusion. Même lorsque les erreurs sont représentées sous forme de diagramme, l'utilisateur doit faire un effort pour interpréter l'information qui lui est présentée et en comprendre l'objet. Il ne conviendrait pas du tout de fournir des données sur les erreurs à un enfant de onze ans qui ne demande qu'un ou deux chiffres. Ce constat nous amène à la question de la quantité de renseignements supplémentaires à fournir.

Certains statisticiens – et quelques utilisateurs avertis – soutiennent que toutes les estimations doivent être accompagnées de renseignements supplémentaires. Certains estiment qu'aucun de ces renseignements supplémentaires ne doivent figurer dans les publications; d'autres relégueraient ces renseignements à un chapitre sur la méthodologie; d'autres encore ne fourniraient le renseignement que sur demande. Comment donc décider de ce qui convient? On peut envisager ce problème de deux façons, en commençant soit par une analyse des produits, soit par une analyse des utilisateurs. Dans le premier cas, on peut répartir les produits en trois groupes :

- ceux qui fournissent des renseignements de base gratuits pour tous, comme les chiffres sur le Canada et l'information générale sur les sites Web;
- les produits et services uniformisés tarifés aux prix du marché, comme un annuaire statistique et les publications sur les tendances sociales;
- le renseignement spécifié par l'utilisateur, comme une tabulation spéciale à partir des vastes ressources en données d'un organisme statistique national.

Une suggestion consisterait à ne fournir aucun renseignement supplémentaire avec les produits du premier groupe, à présenter un chapitre de renseignements supplémentaires pour les produits du deuxième groupe et à publier des renseignements supplémentaires abondants pour les produits du troisième groupe – presque pour chaque estimation produite.

Il faut aussi se demander qui veut des renseignements supplémentaires. Dans le cas de l'enfant de onze ans mentionné précédemment, l'utilisateur n'a besoin d'aucun renseignement supplémentaire. Par contre, un économiste qui tente de modéliser l'économie peut avoir besoin de renseignements supplémentaires spécifiques avec toutes les données fournies. Dans bien des cas, la difficulté pour ce type d'utilisateur est qu'il ne sait pas nécessairement quoi faire de tous ces renseignements supplémentaires.

Que doivent donc faire les méthodologistes et les praticiens? Comme la majorité des utilisateurs ne comprennent pas les renseignements supplémentaires dont nous disposons et qu'ils n'en veulent pas, nous devons poursuivre notre recherche sur les méthodes d'enquête, les sources d'erreur et l'incidence des processus en vue de réduire les nombreuses formes d'erreur et de produire de meilleures estimations. Ainsi, nous réduirons au minimum le besoin de renseignements supplémentaires.

Car tout ce que réclame l'utilisateur moyen, c'est le chiffre qui répond à ses besoins!

RÉFÉRENCES

Farquhar et Farquhar (1891), dans *Economic and Industrial Delusions: A Discourse for the Case of Protection*, New York, Putnam.

Worcester, Sir R. (2004), Observations formulées à la suite d'une conférence présentée au MORI Social Research Institute, septembre 2004.