Une ancienne technique d'échantillonnage :
imparfaite, étonnamment bonne ou optimale?

Par Harry Zvi Davis, Hershey H. Friedman et Jianming Ye

Il y a environ 1 800 ans, un débat a eu lieu concernant la détermination du volume légal d'un œuf, étant donné que le volume varie d'un œuf à l'autre. Les deux acteurs de ce débat ont supposé que la mesure idéale consistait à prendre la moyenne de l'œuf le plus gros et de l'œuf le plus petit.

À première vue, cette mesure est gravement défectueuse. Dans leur ouvrage intitulé *Essential Statistics in Business and Economics,* D. P. Doane et L. E. Seward déclarent qu'elle est facile à calculer, mais qu'il ne s'agit pas d'une mesure robuste de la tendance centrale, parce qu'elle est sensible aux valeurs extrêmes des données. Cependant, si on l'analyse, cette mesure s'avère étonnamment bonne, et, dans l'ancien temps, compte tenu des circonstances, elle pourrait même avoir été optimale.

La *Mishna*, à l'origine une ancienne tradition orale, a été compilée par écrit et éditée il y a environ 1 800 ans par Rabbi Judah Hanasi (le Prince). L'œuf d'une poule a des implications dans les lois traitant de la pureté rituelle, une question importante dans l'ancien temps, puisque les sacrifices et, dans certains cas, la dîme devaient être rituellement purs.

La *Mishna Keilim* (17:6), un traité sur la pureté rituelle, discute de l'établissement du volume légal d'un œuf aux fins de la pureté rituelle. Il ne faut pas oublier qu'il y a 1 800 ans, il n'existait aucune mesure universellement normalisée à laquelle on pouvait se référer. Donc, on pouvait soit relier le volume de l'œuf à une mesure connue, soit fournir une mesure par échantillonnage pour établir le volume.

Selon Rabbi Yehuda, il fallait aller au marché (qui représente un lot d'œufs) et là, choisir « le plus grand des plus grands » œufs et « le plus petit des plus petits » œufs, puis calculer le volume moyen des deux œufs.

Comme il est expliqué dans la *Tosefta* (*Keilim* 2:6:4) — un corpus distinct de règles énoncées par les sages tannaïtiques qui a été compilé à peu près à la même époque que la *Mishna* —, il faut, pour calculer ce volume, placer les deux œufs dans un récipient rempli d'eau à ras bord. Après que l'eau a débordé, les œufs sont remplacés par des objets qui n'absorbent pas l'eau jusqu'à ce que le récipient soit de nouveau rempli à ras bord. Le volume de l'œuf est défini comme étant la moitié du volume des objets placés dans le récipient.

Conceptuellement, Rabbi Yehuda prenait le milieu de l'étendue (moyenne) entre la valeur maximale et la valeur minimale. Rabbi Yosi était fondamentalement d'accord avec la méthodologie de Rabbi Yehuda; cependant, à son avis, le fait que le plus grand œuf et le plus petit œuf sur le marché pourraient ne pas être le plus grand œuf et le plus petit œuf dans la population posait problème. Par conséquent, selon Rabbi Yosi, il fallait utiliser l'œuf médian sur le marché.

Ni Rabbi Yehuda ni Rabbi Yosi n'avaient de doute quant à l'utilisation de deux valeurs extrêmes pour calculer une mesure de la tendance centrale. Cependant, les deux valeurs extrêmes pourraient être un mauvais estimateur, et si la distribution est asymétrique, leur approche donne une estimation biaisée de la moyenne ainsi que de la médiane.

Par souci de cohérence, supposons que la moyenne des volumes des œufs est égale à 4, et que la variance est égale à 1. Pour produire une moyenne de 4 et une variance de 1, si *X* est la variable aléatoire, les variables transformées *Y* qui suivent ont une moyenne de 4 et une variance de 1 : $X \sim \chi^2(6)$ et $Y = 4 + (X - 6)/\sqrt{12}$; $X \sim N(0,1)$ et $Y = X + 4$; $X \sim t(10)$ et $Y = 4 + X/\sqrt{1.125}$; et $X \sim t(3)$ et $Y = 4 + X/\sqrt{3}$. Étant donné un lot de *N* œufs, [...] le biais, la variance et l'erreur quadratique moyenne prévus [diffèrent] lorsque l'on estime la moyenne de la population par l'estimation du milieu de l'étendue [...] pour ces quatre distributions.

Pour une distribution du khi-carré avec 6 degrés de liberté (dl), qui est asymétrique avec étalement vers la droite, l'estimation du milieu de l'étendue est une mesure défectueuse. La mesure est biaisée, et le biais est d'autant plus grand que la taille du lot est grande. Donc, augmenter la taille du lot ne fait qu'accroître l'erreur quadratique moyenne. Pour une distribution *t* avec 3 dl, il n'y a pas de biais, parce que la distribution est symétrique, mais l'erreur quadratique moyenne est beaucoup plus grande que dans le cas de la distribution du khi-carré avec 6 dl. Cela reviendrait à échantillonner la personne la plus lourde et la personne la plus légère dans un grand groupe d'hommes adultes. Quelqu'un vivant dans une collectivité où vit la personne la plus lourde — Manuel Uribe de Mexico a pesé à un moment donné 1 320 livres — obtiendra en prenant la moyenne de l'observation pour la personne la plus lourde et de l'observation pour la personne la plus légère, un poids moyen de plus de 660 livres. Plus la taille du lot augmente, moins bon devient l'estimateur.

Distribution normale

Mais considérons une distribution normale. Même pour un petit lot de 20, il n'y a pas de biais et l'erreur quadratique moyenne n'est que de 0,14. À mesure que la taille du lot augmente, la variance diminue. Si l'on analyse une distribution *t* avec 10 dl (qui comporte de nombreuses valeurs aberrantes), l'estimation du milieu de l'étendue produit une erreur quadratique moyenne d'environ 0,3. Même si la variance augmente quand la taille du lot augmente, l'erreur quadratique moyenne varie peu.

Selon les auteurs de l'article intitulé « The Influence of Body Size, Breeding Experience, and Environmental Variability on Egg Size in the Northern Fulmar », publié dans le *Journal of Zoology*, empiriquement, le volume de l'œuf suit une loi normale. En outre, comme l'ont souligné C. J. Adams et D. D. Bell dans leur article publié dans le *Journal of Applied Poultry Research*, les chercheurs supposent généralement que la taille et le poids des œufs suivent une loi normale et font appel à la régression et à la corrélation pour déterminer les facteurs qui influent sur le poids de l'œuf.

En supposant que le volume des œufs suit une loi normale, nous comparons l'efficacité de l'utilisation de l'estimation du milieu de l'étendue par opposition à la moyenne d'échantillons aléatoires. Nous présentons ensuite l'échantillon équivalent à un échantillon aléatoire (sans remise) ayant la même exactitude que l'estimation du milieu de l'étendue [...]. Bien que deux œufs seulement soient utilisés pour calculer le milieu de l'étendue, si le lot (marché) contient 100 œufs, cela équivaut à un échantillon aléatoire de 11 œufs. Si la taille du lot est de 1 000, l'estimation du milieu de l'étendue équivaut à un échantillon aléatoire de taille 16.

Sensibilité

Dans quelle mesure l'erreur quadratique moyenne est-elle sensible au choix de l'œuf maximal et de l'œuf minimal? Qu'arrive-t-il si une légère erreur est commise lors du choix de l'œuf maximal et de l'œuf minimal et qu'un très gros œuf et un très petit œuf sont choisis à la place? Pour simuler le processus, nous tirons un sous-échantillon de 10 % des plus gros œufs dans le

lot et choisissons aléatoirement un œuf dans le sous-échantillon comme étant l'œuf très gros. De même, nous tirons un sous-échantillon de 10 % des plus petits œufs dans le lot et choisissons aléatoirement dans le sous-échantillon un œuf comme étant l'œuf très petit. [...]

Bien que la variance augmente légèrement pour des tailles de lot plus grandes, les résultats sont presque identiques à ceux obtenus en utilisant le plus gros œuf et le plus petit œuf.

Mesure des erreurs dans le calcul de la moyenne

Mais un autre élément doit être pris en considération. Pour un échantillon aléatoire de 16, par exemple, le volume d'eau déplacé doit être divisé en 16 parties égales. Conceptuellement, il s'agit d'un exercice trivial. Cependant, comme peut en témoigner tout parent qui a essayé de diviser également une portion d'aliment entre des enfants qui se chamaillent, opérationnellement, la tâche n'est pas triviale. En outre, plus le nombre de parties en lesquelles le volume d'eau déplacé doit être divisé est élevé, plus l'erreur de mesure est grande. Donc, tout gain d'efficacité dans l'estimation de la moyenne de population réalisé en augmentant le nombre d'œufs dans l'échantillon pourrait être plus que qu'annulé par l'erreur de mesure associée à la division du résultat en parties égales. Il se pourrait que, puisque la méthode du milieu de l'étendue donne des résultats supérieurs à ceux de tout échantillon de taille supérieure à 16, utiliser la méthode du milieu de l'étendue (qui ne nécessite qu'une division en deux parties égales) puisse produire une meilleure estimation du volume de l'œuf.

Optimalité

Est-il possible de trouver une technique d'échantillonnage différente, qui serait meilleure que la méthode du milieu de l'étendue? Une option consiste à utiliser les deux œufs les plus gros et les deux œufs les plus petits. Cela réduit la variance d'un tiers [...]. Cependant, les œufs doivent maintenant être divisés en quatre parties, au lieu de deux. Il se peut que l'accroissement de l'erreur de mesure de la moyenne d'échantillon surpasse la diminution de la variance d'échantillon. Mais considérons l'utilisation du deuxième plus grand et du deuxième plus petit œuf. Particulièrement pour les grandes tailles de lot, cela réduit la variance d'environ la moitié. Donc, si l'objectif est d'utiliser un échantillon de deux œufs pour estimer la taille moyenne des œufs dans la population, l'utilisation du deuxième plus grand et du deuxième plus petit œuf est la meilleure option examinée ici. Pour une taille de lot de 1 000, l'échantillon équivaut à choisir plus de 33 œufs aléatoirement [...].

[Extraits de l'article intitulé « An Ancient Sampling Technique: Flawed, Surprisingly Good, or Optimal? » publié dans *Chance*, vol. 24, n^{o} 1, 2011]