Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

LA CONFIANCE EN LA TAILLE D'ÉCHANTILLONNAGE OPTIMALE FONDÉE SUR DES DONNÉES ANTÉRIEURES

Charles Fleming et Richard McGuinness[1]

RÉSUMÉ

Dans le cas d'une enquête à plusieurs variables fondée sur un échantillonnage aléatoire simple, le calcul de la taille optimale d'échantillon revient à résoudre un problème de programmation stochastique où chaque contrainte correspond à une estimation bornée de la variance d'une estimation donnée. Le problème est stochastique parce que l'ensemble de données recueillies lors d'une enquête antérieure fait que les composantes de chaque contrainte sont des variables aléatoires; par conséquent, la taille calculée de l'échantillon est elle-même une variable aléatoire qui dépend de la qualité de l'ensemble de données. Le recours à une méthode de Monte Carlo permet de produire une loi de distribution empirique de la taille optimale d'échantillon en vue de déterminer la probabilité d'obtenir la précision voulue. À chaque ensemble de données recueillies antérieurement correspondent une taille d'échantillon et une répartition entre strates optimales. L'examen de ces caractéristiques sur plusieurs périodes consécutives permet de repérer les strates problématiques et de déceler une tendance quant à la stabilité des données. Il pourrait révéler une courbe de la taille d'échantillon en fonction du temps de nature oscillatoire due à la dépendance d'une répartition à l'égard d'une autre.

MOTS CLÉS : Taille optimale d'échantillon; programmation convexe; plan de sondage.

1. FORMULATION

Lorsque l'obtention de données statistiques dépend d'une collaboration volontaire, l'incertitude entourant la réponse à une question donnée et la variabilité prévue des données recueillies jouent un rôle important dans la détermination de la taille de l'échantillon. Le problème consistant à préciser la taille appropriée d'un échantillon peut être décrit entièrement sous forme d'un énoncé mathématique. Toutefois, aussi rigoureux que puisse paraître cet énoncé, la détermination de la taille optimale d'un échantillon est un problème stochastique et non un problème déterministe. En fait, la taille optimale d'un échantillon est une variable aléatoire qui regroupe en un seul nombre les caractéristiques aléatoires des estimations, celles des estimations de la variance et celles de l'estimation de la coopération des personnes sélectionnées pour participer à l'enquête. Puisque la taille optimale d'échantillon est une fonction de ces composantes importantes, sa loi de distribution peut , d'une part, servir au choix d'une taille d'échantillon en fonction des intervalles de confiance et, d'autre part, être utilisée dans un programme de contrôle de la qualité.

Afin de rendre une enquête plus efficace, un statisticien pourrait décider de recueillir des renseignements sur plusieurs sujets en une seule interview au lieu de les interviewer individuellement. Donc, dans le cas d'une enquête à plusieurs variables, la taille de l'échantillon doit tenir compte simultanément des composantes de l'enquête pour toutes les variables. Le statisticien qui connaît a priori certains résultats d'une enquête peut concevoir une enquête plus efficace. À cet égard, la stratification de la liste des unités que compte une population est une pratique courante. Une autre consiste à utiliser l'information provenant d'une enquête antérieure pour décrire les caractéristiques de la population courante, à condition qu'il soit raisonnable de supposer que la population courante ne diffère pas trop de la population antérieure. Lors de la conception d'une enquête, l'un des grands objectifs consiste à recourir à ces pratiques afin de produire des estimations précises au meilleur prix possible.

[1]Charles Fleming et Richard McGuinness, National Agricultural Statistics Service, U.S. Department of Agriculture, 1400 Independence Avenue, SW, Washington, DC 20250, USA.

Les méthodes telles que la stratification de la liste des unités d'une population facilitent la réalisation de cet objectif, mais la précision des estimations est extrêmement sensible à l'existence sur la liste d'unités classées incorrectement. De surcroît, le problème de la définition d'une stratification optimale n'a pas encore été résolu dans le cas des enquêtes à plusieurs variables. Néanmoins, étant donné une stratification particulière, la détermination de la taille optimale d'échantillon peut être formulée comme un problème de programmation convexe, de la façon suivante :

$$\min \sum_{h\ 0\ strata} c_h n_h$$

$$de\ sorte\ que$$

$$\sum N_h(N_h - \hat{\rho}_h n_h)\frac{s_h^2}{n_h} \leq (CV)^2$$

$$0 \leq n_h \leq N_h$$

où la somme est calculée sur toutes les strates, c_h représente le coût de la réalisation de l'enquête dans la strate h, n_h représente la taille optimale d'échantillon pour la strate h, $\hat{\rho}_h$ représente l'estimation du taux de réponse des personnes qui sont sélectionnées pour participer à l'enquête et $\hat{\tau}$ représente l'estimation de la la quantité d'un produit. Dans le cas d'un problème multivarié, une contrainte est définie pour chaque produit. Si l'on donne à n_h une valeur suffisamment grande, la précision estimée de $\hat{\tau}$ augmentera jusqu'à ce que le coefficient de variation tombe au dessous d'une valeur précisée, mais au prix de l'augmentation du coût de l'enquête. Ce problème de programmation convexe est, en fait, un problème de programmation stochastique, car il contient des variables aléatoires; par conséquent, sa solution est une variable aléatoire à laquelle est associée une loi de distribution que nous utiliserons pour évaluer la qualité des données recueillies antérieurement.

L'idée fondamentale selon laquelle la variabilité naturelle des données chronologiques influe sur la fiabilité de la valeur calculée de la taille optimale d'échantillon ne devrait jamais être perdue de vue. Savoir dans quelle mesure la variation aléatoire des données influe sur la taille optimale d'un échantillon offre deux avantages importants. En premier lieu, ce renseignement permet au statisticien de construire une loi de distribution de la taille optimale d'échantillon de sorte qu'il puisse choisir la taille d'échantillon d'après l'examen des intervalles de confiance. Autrement dit, connaître la loi de distribution de la taille optimale permet d'estimer la probabilité que l'enquête produira des estimations ayant la précision voulue pour une taille donnée d'échantillon. En deuxième lieu, le renseignement permet au statisticien d'évaluer l'efficacité de la méthode choisie pour stratifier la liste grâce à l'examen de la cohérence de la taille optimale dans chaque strate au fil du temps.

Dans le cas d'une enquête à plusieurs variables, seules les méthodes numériques permettent de résoudre le problème de programmation convexe formulé plus haut. Nous avons utilisé ces mêmes techniques numériques dans un étudede Monte Carlo pour produire plusieurs centaines de tailles optimales par strate, de façon telle que chaque calcul provienne d'un ensemble de variables représentant l'erreur-type, la non-réponse et le total étendu à la population. L'ensemble total de nombres produits grâce à ces calculs a permis de créer une loi de distribution empirique de la taille optimale d'échantillon.

2. DISTRIBUTION EMPIRIQUE

Avant de procéder à la simulation de Monte Carlo en vue de déterminer la taille optimale d'échantillon, il faut vérifier s'il existe une corrélation entre le taux de réponse des personnes sélectionnées pour participer à l'enquête, les variances d'échantillonnage et les estimations pour s'assurer que l'hypothèse théorique voulant que ces éléments soient indépendants est satisfaite. Pour déterminer s'il existe une corrélation, nous avons produit des graphiques semblables à celui présenté à la figure 1. Cette figure, produite d'après les valeurs

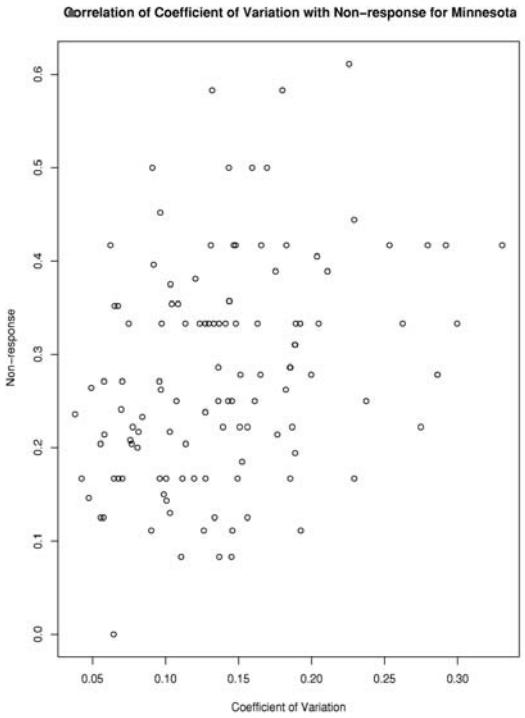

Figure 1. Il n'existe aucune corrélation apparente entre le c.v. et la non-réponse.

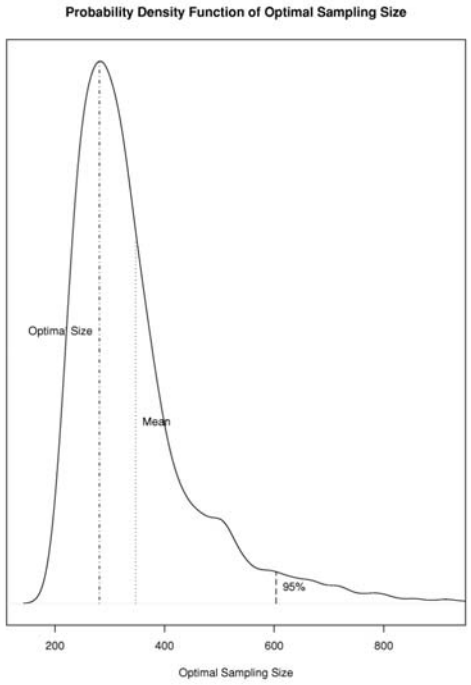

Figure 2. . Selon la loi de distribution empirique, le 95^{e} centile de l'obtention du c.v. requis correspond à une taille d'échantillon de 603.

calculées pour la Quarterly Agricultural Labor Survey pour un ensemble de treize trimestres et onze strates, montre qu'il n'existe aucune corrélation apparente entre le coefficient de variation (c.v.) et la non-réponse. Après examen de tous les autres graphiques possibles, nous avons appliqué la méthode de Monte Carlo en considérant la variance d'échantillonnage, le total de population et la non-réponse comme étant indépendants.

Nous ne connaissons pas les distributions exactes de la variance d'échantillonnage, de la non-réponse et du total étendu à la population, mais nous avons le choix entre plusieurs bonnes approximations. Par exemple, nous supposons que le carré de l'erreur-type, s^2, obéit à une loi de distribution X^2. En outre, comme l'étendue de la non-réponse est limitée à l'intervalle [0,1], nous supposons que la non-réponse obéit à une loi de distribution bêta. Afin de pouvoir concevoir un étude de Monte Carlo stable où les variables aléatoires produites sont toutes non négatives et ont un comportement raisonnable, nous avons utilisé une loi de distribution bêta plutôt qu'une loi de distribution normale pour décrire le total de population.

Étant donné la complexité du problème, il n'est pas possible de calculer une loi de distribution théorique de la taille d'échantillon optimale. Par conséquent, nous avons obtenu, par une méthode de Monte Carlo, une loi de distribution empirique pour l'une des Quarterly Agricultural Labor Surveys réalisée régulièrement. La fonction de densité de probabilité empirique produite d'après l'ensemble de 1 000 tailles optimales calculées en se servant de valeurs générées aléatoirement de la variance d'échantillonnage, de la non-réponse et du total de population est illustrée à la figure 2. Sur cette figure, nous remarquons surtout la forme asymétrique de la distribution et, avant tout et par dessus tout, la queue manifestement lourde à droite. Ces deux caractéristiques laissent entendre qu'il faudra, pour obtenir le coefficient de variation souhaité, utiliser une taille d'échantillon plus grande que ne le laisse supposer le problème déterministe.

D'après la distribution empirique, nous constatons que la taille d'échantillon nécessaire pour obtenir le coefficient de variation requis pour cette enquête particulière à un niveau de confiance de 95 % est égale à 603. Par ailleurs, la probabilité d'obtenir le c.v. requis pour la taille d'échantillon de 281 établie par la méthode probabiliste, probabilité qui est représentée par la surface sous la courbe, n'est que de 35 %. Même pour la taille d'échantillon simulée de 337, la probabilité d'obtenir le c.v. requis n'est que de 60 %, autrement dit à peine plus élevée que si l'on jouait à pile ou face.

Pour déterminer les causes de la queue lourde, nous devons examiner certains résultats détaillés. Le tableau 1 présente les résultats de la simulation de Monte Carlo pour l'Agricultural Labor Survey réalisée au Minnesota. Nous constatons que, à cause de la queue lourde de la distribution, la taille optimale simulée est plus grande que la taille optimale déterministe pour chaque strate sauf une. Notre statistique CHI2, qui figure dans la colonne intitulé CHI2 , est analogue à la statistique X^2 de qualité d'ajustement. Cette statistique spéciale mesure l'influence de la lourde queue de droite sur la taille d'échantillon par sommation des quotients du carré de l'écart entre les tailles optimales d'échantillon déterministe et simulée, par la taille déterministe. Il s'agit d'une mesure de la qualité des données recueillies antérieurement qui est reflétée dans la loi de distribution empirique par la queue lourde à droite.

Tableau 1.

Strate	Taille de la population	Estimation Non-réponse	Taille optimale	Taille optimale simulée	CHI2
50	9 362	0,278	5	6,2	0,28
55	12 423	0,167	9	12,78	1,59
70	5 020	0,286	10	12,44	0,59
75	612	0,250	1	1,54	0,30
79	28 668	0,146	42	61,65	9,38
85	787	0,354	10	11,28	0,16
92	16 198	0,167	81	99,05	4,02
93	2 001	0,167	31	34,63	0,42
94	834	0,250	40	45,26	0,69
95	245	0,333	29	32,41	0,40
96	23	0,250	23	19,45	0,54
Total			281	336,95	18,42

Le tableau 2 donne les valeurs de CHI2 pour treize enquêtes consécutives réalisées sur une période de quatre ans, de la plus récente, représentée par le chiffre 13, à la plus ancienne, représentée par le chiffre 1. Lorsqu'on parcours le tableau, les valeurs élevées de CHI2 obtenues pour les enquêtes 6 et 2 sautent aux yeux et font penser qu'il faut en rechercher la cause. Nous avons découvert, lors de l'examen des circonstances de l'enquête 6, que la taille réelle d'échantillon était trop petite. De toute évidence, encouragés par la précision satisfaisante des estimations lors des enquêtes antérieures, les statisticiens ont essayé d'utiliser un échantillon de plus petite taille. Au lieu de sélectionner l'échantillon habituel de 350 unités, ils ont utilisé un échantillon de 318 unités. S'ils avaient pu, à l'époque, utiliser une méthode de programmation stochastique, ils auraient constaté que la qualité des données de l'enquête antérieure était, en fait, problématique et qu'il aurait fallu utiliser un échantillon de plus grande taille au lieu d'un échantillon plus petit. En réponse à cette expérience, nous observons une diminution générale de la valeur de CHI2 sur l'ensemble des strates au fil du temps jusqu'à l'enquête 13, qui est la plus récente, grâce à une meilleure stratification de la liste et à l'utilisation d'échantillons de plus grande taille.

Une tendance se dégage non seulement au fil du temps, mais aussi en fonction de la strate. La valeur de CHI2 est systématiquement élevée pour les strates 55, 79 et 92, apparemment à cause de la classification de certains éléments de la liste dans des strates inappropriées. Nous observons une diminution régulière de la valeur de CHI2 dans toutes les strates après l'enquête 8 jusqu'à l'heure actuelle, due à l'adoption de nouvelles définitions

des strates et à l'utilisation d'échantillons de plus grande taille. Pareillement, pour les strates 50, 60 et 96, la valeur de CHI2 est systématiquement faible, indiquant que la répartition de l'échantillon pour ces strates est inadéquate mais que les définitions de strates sont bonnes. Ces strates fournissent généralement de bonnes estimations et peu de valeurs aberrantes.

Ni l'examen de l'asymétrie ni celui de l'aplatissement n'est très utile. Alors que la mise en tableau des valeurs de la statistique CHI2 permet de déceler des tendances explicables lorsque l'on examine les données en fonction du temps et de la strate, celle des statistiques d'asymétrie et d'aplatissement de la loi de distribution empirique ne permet de discerner aucune tendance, si ce n'est que ces valeurs sont presque toujours négatives pour la strate 96 et fortement positive pour les autres.

Tableau 2.

Strate	13	12	11	10	9	8	7
50	0,28	10,35	1,41	0,92	1,24	0,01	2,45
55	1,59	3,62	0,23	20,42	10,21	34,43	4,52
70	0,59	2,61	0,22	9,11	3,45	15,28	1,32
75	0,3	0	0	0,82	0,85	0,46	0,03
79	9,38	6,34	0,54	0,46	2,58	0,31	0,44
85	0,16	1,05	0	3,24	0,61	2,67	0,54
92	4,02	7,4	0,3	25,81	5,92	105,43	32,03
93	0,42	2,57	0,06	10,61	3,75	20,23	1,27
94	0,69	1,14	0,16	8,83	4,65	16,16	4,57
95	0,4	1,34	0	0,96	0,83	1,86	0,51
96	0,54	0,02	0,01	0,54	0,66	0,91	0,23

Strate	6	5	4	3	2	1	
50	0	0	0	0	27,44	3,6	
55	28 353,95	14,22	0,14	0,47	6569	4,28	
70	8 508,47	25,57	1,1	0,47	16,1	6,31	
75	870,25	0,12	0,34	0,02	32,05	1,3	
79	49 508,01	115,92	1,65	0,92	192,57	17,18	
85	135,15	19,34	0,92	0,11	46,06	2,39	
92	23 800,76	20,16	1,15	0,89	0,23	9,73	
93	1 695,78	4,63	0,42	0,27	901,41	2,33	
94	157,3	3,92	0,1	0,9	2378,55	0,37	

95	15,48	5,26	0,16	0,11	8,35	0,94	
96	0	0,74	0,41	0,24	1,08	0,17	

3. TAILLE DE L'ÉCHANTILLON À L'ÉTAT D'ÉQUILIBRE

Lors de l'examen de l'enquête 6, nous avons remarqué la corrélation manifeste entre la taille de l'échantillon de l'enquête courante et celle de l'échantillon de l'enquête précédente. La figure 3 illustre cette corrélation. Nous avons tracé sur le graphique la droite de meilleur ajustement par les moindres carrés qui montre plus clairement la relation linéaire.

La droite montre qu'il existe une relation inhérente entre la taille optimale d'échantillon calculée et la taille réelle de l'échantillon utilisé lors de l'enquête précédente. Cette relation n'est pas étonnante, car la taille de l'échantillon de l'enquête précédente influe directement sur la variabilité des données recueillies lors de l'enquête courante, données qui deviendront en bout de ligne les données chronologiques utilisées pour

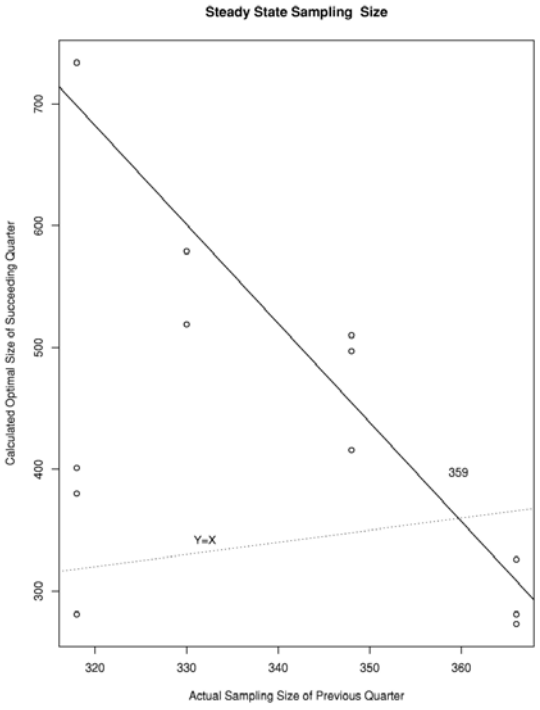

Figure 3. Relation entre la taille optimale de l'échantillon calculée et la taille actuelle d'échantillon utilisée lors de l'enquête précédente.

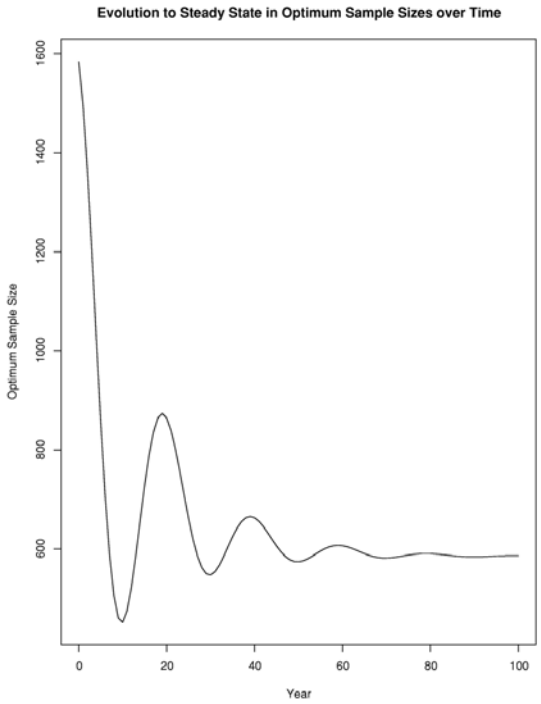

Figure 4. Illustration qualitative de l'évolution au fil du temps de la taille optimale d'échantillon vers une taille d'échantillon d'état d'équilibre.

déterminer la taille de l'échantillon de l'enquête suivante. Autrement dit, l'effet d'une taille choisie d'échantillon se propage lors des enquêtes subséquentes comme dans un processus markovien. Un échantillon de grande taille lors d'une enquête donnée pourrait induire l'utilisation d'un échantillon de plus petite taille lors de l'enquête suivante, parce qu'il a produit de bonnes données de haute précision, tandis qu'inversement, un échantillon de petite taille pourrait induire l'utilisation d'un échantillon de plus grande taille lors de l'enquête suivante, parce qu'il a produit des données

bruitées de faible précision. Ainsi, la taille optimale d'un échantillon dépend de la taille de l'échantillon de l'enquête précédente. Donc, les tailles d'échantillon ne sont pas indépendantes d'une enquête à l'autre. Par conséquent, la courbe représentant l'évolution de la taille de l'échantillon d'une enquête à l'autre pourrait être de nature oscillatoire, comme celle représentée qualitativement à la figure 4.

En dernière analyse, en procédant par tâtonnement, on atteindra une taille d'échantillon d'état d'équilibre correspondant à la valeur de convergence de la taille optimale d'échantillon calculée et de la taille réelle de l'échantillon. En reconnaissant qu'il existe une relation inverse entre une taille d'échantillon donnée et la taille optimale d'échantillon calculée pour l'enquête suivante, nous pouvons affirmer que la taille optimale d'échantillon d'état d'équilibre sera celle pour laquelle les tailles d'échantillon de deux enquêtes consécutives sont égales. Au lieu d'utiliser la méthode par tâtonnement sur un grand nombre d'années pour déterminer la taille d'échantillon à l'état d'équilibre, nous pouvons déterminer graphiquement cette taille en relevant le point d'intersection de la droite en pointillés Y = X avec la courbe représentant la dépendance entre la taille optimale d'échantillon et les données historiques. Ces droites sécantes sont représentées à la figure 3 dans le cas du Minnesota, pour lequel la taille optimale d'échantillon à l'équilibre est égale à 359.

Une autre méthode de détermination de la taille d'échantillon à l'état d'équilibre consiste à agréger les variances calculées pour les enquêtes antérieures, puis à utiliser cette variance agrégée pour calculer une taille optimale d'échantillon. Dans le cas d'une enquête à une seule variable, pour laquelle les estimations, les variances d'échantillon, les taux de réponse et la stratification de la liste sont assez stables au fil du temps, la méthode d'agrégation des variances pourrait représenter une solution simple. Par contre, pour une enquête à plusieurs variables, surtout si les composantes susmentionnées ne sont pas stables, le groupement des renseignements sur une enquête peut se faire de façon transparente en résolvant le problème de programmation convexe, afin de pouvoir obtenir directement une taille d'échantillon à l'état d'équilibre, comme nous l'avons montré plus haut.

D'après la relation fonctionnelle entre la taille d'échantillon et les estimations, les variances d'échantillonnage, les taux de réponse et la stratification donnée dans la formulation d'une programmation convexe, il est possible d'obtenir une taille optimale d'échantillon à l'état d'équilibre. Comme cette dernière est une variable aléatoire, sa loi de distribution permet non seulement de déterminer la probabilité d'obtenir la précision requise d'une estimation pour une taille d'échantillon donnée, mais aussi, grâce à l'examen de sa forme asymétrique et de sa queue lourde à droite au moyen d'une statistique comme CHI2, de surveiller la qualité des données d'une strate à l'autre et au fil du temps.