REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

Quelques points se rapportant aux calculs de variance des sondages

Revue de statistique appliquée, tome 13, n° 4 (1965), p. 45-50

<http://www.numdam.org/item?id=RSA_1965__13_4_45_0>

QUELQUES POINTS SE RAPPORTANT AUX CALCULS DE VARIANCE DES SONDAGES (1)

P. THIONET

Professeur à la Faculté des Sciences de Poitiers

1 - INTRODUCTION

Par calculs de variance on entend les procédés par lesquels on évalue l'importance des erreurs d'échantillonnage en utilisant les seules données échantillon. Le but essentiel d'un sondage est d'obtenir les estimations de certaines grandeurs (économiques, démographiques, etc), et le calcul des variances de ces estimations ne saurait être une fin en soi. Aussi le plus souvent, au moins en France, ces calculs sont-ils très approximatifs ; on ne les fait d'ailleurs que pour une part assez limitée des grandeurs estimées sur échantillon (et on a bien raison de ne pas s'attarder lorsqu'il s'agit d'appliquer des formules inextricables).

Pourtant il n'est pas sans intérêt d'obtenir (par calcul de variance) des indications sur les erreurs affectant les estimations et imputables au fait qu'on dispose d'un échantillon de données au lieu de collecter la totalité des informations. Souvent ces erreurs d'échantillonnage sont notablement moins grandes que les erreurs dues aux méthodes de relevé ou d'enregistrement ; ce qui permettrait de chercher dans une réduction de la taille des échantillons et une amélioration des techniques d'observation une meilleure utilisation des ressources disponibles.

On doit déplorer qu'un calcul de variance exécutable à peu de frais soit omis ; mais il est fréquent que des calculs de variance omis soient impossibles ou encore semblent longs, difficiles et (pour tout dire) onéreux ; ils n'intéressent finalement qu'un nombre assez restreint de spécialistes, alors que, pour le même prix, des tabulations plus détaillées et entrecroisées peuvent fournir des informations intéressant un public bien plus large (même si ces informations se trouvent en fait erronées parce qu'entachées d'énormes erreurs d'échantillonnage).

Il pourrait se faire qu'on évitât des calculs de variance, de peur qu'ils ne révélassent des erreurs d'échantillonnage tellement grandes que le sondage les concernant cesserait d'intéresser les personnes trop scrupuleuses. Une attitude également condamnable consisterait à effectuer des calculs de variance "tranquillisants", en faisant toutes les approxi-

---

(1) Notre attention a été attirée sur quelques points de la théorie des sondages à l'occasion d'une réunion d'experts Statisticiens des Nations-Unies à Genève (1.12 juillet 1963). La présente note est le début d'une série d'articles sur ce sujet. Que M. Loftus, Directeur du Bureau de Statistique de l'O.N.U. trouve ici l'expression de notre gratitude.

mations par défaut. Il n'est d'ailleurs pas toujours possible de faire (à défaut de véritable calcul de variance) le calcul de deux limites (supérieure et inférieure) d'une variance, ces limites restant (disons) dans le rapport de 1 à 2 ; ce serait pourtant une solution réaliste.

2 - INSUFFISANCES DE PROCEDES ACTUELS

Les procédés actuels de calcul de variance laissent subsister des lacunes. D'une part le calcul de variance d'un *sondage systématique* en est au stade de l'exploration (alors que le sondage systématique est le plus utilisé des procédés de sondage). Il en est de même du sondage *en grappe* et de ce qui s'y rattache.

En second lieu le calcul de variance est souvent anormalement grossier et on ne peut avoir qu'une idée très vague d'une variance quand on connait son estimation.

Enfin certaines méthodes (théoriquement intéressantes) conduisent parfois à des estimations *négatives* de variances positives, et de ce fait sont inutilisables.

Nous ne sommes hélas pas en mesure de faire face ici à tous ces problèmes, mais il convient d'en prendre conscience.

3 - UN CAS DE CALCUL DE VARIANCE TROP GROSSIER

Soit un échantillon $\mathcal{S}$ de grande taille comme c'est le cas habituellement dans les sondages ; il sert à estimer $\overline{X}$ (moyenne de la population $\mathcal{U}$) par $\overline{x}$ (moyennes échantillon), et aussi à estimer la variance $v\overline{x}$. Celle-ci a une expression théorique, fonction de paramètres de $\mathcal{U}$.

Si l'on estime les paramètres de $\mathcal{U}$ à l'aide de $\mathcal{S}$ et qu'on porte ces estimations dans $\mathcal{V}$, on obtient une estimation V de $v\overline{x}$. Si l'on a pu employer une fonction linéaire et des estimations sans biais des paramètres, on arrive à une estimation V *sans biais*. On n'est pas à l'abri d'une estimation finale négative V ; et pourtant cette méthode de calcul pas à pas est relativement précise.

Une méthode beaucoup plus rapide consiste à faire une dichotomie de l'échantillon $\mathcal{S}$ et à en tirer 2 estimations *indépendantes* $\overline{x}_1$ et $\overline{x}_2$ de $\overline{X}$, alors on peut toujours estimer $v\overline{x}$ sans biais par

$$\frac{1}{4}(\overline{x}_1 - \overline{x}_2)^2 = v \quad ; \text{ avec } \quad \overline{x} = \frac{\overline{x}_1 + \overline{x}_2}{2}$$

Il est clair que v ne sera jamais négatif ; mais on s'expose en revanche à ce que cet estimateur soit aberrant.

Comme l'échantillon est grand, $\overline{x}_1$ $\overline{x}_2$ $\overline{x}$ ont des distributions (pratiquement) de Laplace-Gauss ; $\overline{x}_1$ et $\overline{x}_2$ sont 2 valeurs indépendantes d'une certaine variable de Laplace-Gauss. On sait que $(x_1 - x_2)^2/2$ est estimation sans biais de la variance σ^2 de cette variable. Enfin, on a

$$v\overline{x} = v\left(\frac{\overline{x}_1 + \overline{x}_2}{2}\right) = \frac{1}{4}\left[v\overline{x}_1 + v\overline{x}_2 + 2\ \text{Cov.}\ (\overline{x}_1.\overline{x}_2)\right]$$

$\Longrightarrow$ est $v\bar{x} = \frac{1}{2}$ est σ^2, car Cov.$(\bar{x}_1\ \bar{x}_2) = 0$

$\Longrightarrow$ est $v\bar{x} = \frac{1}{4}(\bar{x}_1 - \bar{x}_2)^2$

On sait aussi que $(\bar{x}_1 - \bar{x}_2)^2/2$ est un $\sigma^2\ \chi^2$ à 1 degré de liberté : car $(\bar{x}_1 - \bar{x}_2)$ est une variable de Laplace Gauss de variance $2\sigma^2$.

Par suite cette estimation a 10 chances sur 100 d'être supérieure à 2,7 fois σ^2.

et 30 chances sur 100 d'être inférieure à 0,15 fois σ^2.

et même 20 chances sur 100 d'être inférieure à 0,06 fois σ^2).

Probabilité $(v < 0{,}46\ v\bar{x}) = 50\ \%$

Probabilité $(v > 1{,}64\ v\bar{x}) = 20\ \%$

C'est ce que nous considérons comme un calcul de variance sans portée pratique, parce que l'estimation de la variance a toutes chances d'être aberrante.

Un procédé meilleur consiste à scinder l'échantillon $\mathcal{S}$ disons en 10 sous-échantillons de taille encore raisonnablement grande et à calculer un $\sigma^2\ \chi^2$ avec *9 degrés* de liberté. Alors l'estimation de σ^2 n'a plus que :

10 chances sur 100 d'être inférieure à $\frac{1}{2}\sigma^2$

4 chances sur 100 d'être supérieure à $2\sigma^2$.

Un tel calcul de variance présenterait donc un réel intérêt pratique. Mais le cas où il est praticable sont évidemment moins étendus que ceux où le premier procédé s'applique, encore qu'on doive dans le cas le plus général renoncer à l'un et à l'autre. Comme des erreurs ont pu être commises à ce sujet, il convient de s'étendre un peu.

4 - CONDITIONS DANS LESQUELLES DES PROCEDES S'APPLIQUENT - OU NE S'APPLIQUENT PAS

Personne ne saurait imaginer que, l'échantillon $\mathcal{S}$ étant représenté par 3 000 cartes perforées, les estimations $\bar{x}_1$ et $\bar{x}_2$ s'obtiennent respectivement sur les cartes n° 1 à 1500 d'une part, n° 1501 à 3000 d'autre part ; il va de soi que les 2 lots peuvent représenter l'un le nord et l'autre le midi de la France, et non pas l'un et l'autre la totalité de la France. En revanche on pourrait croire que, si $\bar{x}_1$ est la moyenne des x figurant sur les cartes dont le numéro est impair et $\bar{x}_2$ celle des cartes de numéros pairs, on se trouve dans les conditions prévues par le calcul. Or, ce n'est pas en général correct.

Si l'échantillon a été *tiré au sort à 1 seul degré*, même avec stratifications (fractions de sondage égales dans les strates) et si les cartes d'une même strate portent des numéros consécutifs, la technique "*pair-impair*" est valable (en négligeant l'erreur due aux strates représentées

par un nombre *impair* de cartes). Si les strates ont des fractions sondées différentes, donc des poids différents dans $\bar{x}$, il suffit de conserver ces poids dans $\bar{x}_1$ et $\bar{x}_2$.

En revanche si l'échantillon est prélevé de façon *systématique* (= en progression arithmétique) ou encore s'il provient de tirages au sort *à plusieurs degrés*, le procédé est incorrect.

En effet (comme on vient de le voir au § 3) le calcul suppose

$$\text{covariance } (\bar{x}_1, \bar{x}_2) = 0$$

autrement dit :

$$\text{corrélation } (\bar{x}_1, \bar{x}_2) = 0$$

a) Si $\bar{x}_1$ et $\bar{x}_2$ sont les moyennes de 2 échantillons systématiques avec la fraction sondée f/2, il n'y a aucune raison pour que cov $(\bar{x}_1 \bar{x}_2) = 0$. Et même si les 2 sous-échantillons ne sont pas systématiques, ils nous renseigneront sur l'échantillon $\mathcal{S}$ (systématique) et non sur la population $\mathcal{U}$.

b) De même si $\mathcal{S}$ est tiré à 2 degrés et si $\bar{x}_1$ et $\bar{x}_2$ sont calculés chacun sur (disons) la moitié des ménages-échantillons de chaque commune-échantillon, il est clair que $(\bar{x}_1 - \bar{x}_2)^2$ ne peut nous informer sur la variance $v\bar{x}$ mais seulement *sur la variance entre ménages à l'intérieur des communes échantillon*. Si on décompose $v\bar{x}$ comme d'habitude en :

$$v\bar{x} = v_i + v_e$$

$(\bar{x}_1 - \bar{x}_2)^2$ ne nous informe que sur la variance interne v_i et non sur la variance externe v_e.

Si nous disposons toutefois de 2 communes (u ps)-échantillon par strate, on peut constituer les 2 sous-échantillons de ces strates avec les ménages des communes différentes. Bref, on peut alors procéder à une dichotomie de l'échantillon telle qu'on ait : corrélation $(x_1, x_2) = 0$. Mais il va de soi que la scission de $\mathcal{S}$ en (disons) 10 sous-échantillons, procurant $\bar{x}_1$ $\bar{x}_2$ ... $\bar{x}_{10}$, grâce à 10 communes tirées de chaque strate, n'est pas réaliste :

On ne va pas construire un plan de sondage comprenant des strates aussi grandes et aussi peu nombreuses, avec pour seul motif la recherche d'un calcul d'erreur facile.

Ce qu'on ferait peut-être (et avec de bonnes raisons), c'est un regroupement a posteriori de strates comparables, de façon à constituer *artificiellement* 10 sous-échantillons. Par exemple on trouverait bien dans l'échantillon effectivement soumis à l'enquête 10 communes rurales de l'Ouest, 10 petites villes de l'Est, etc... Mais les nombres de ménages sondés dans chacune des 10 unités primaires ainsi rapprochées ne coïncideraient pas, ce qui sera la cause de difficultés supplémentaires.

5 - CALCUL INCORRECT UTILISABLE

Quand on évalue la variance par la formule $\sum_1^{10} (\bar{x}_i - \bar{x})^2 / 9 = V_o$ on obtient donc un résultat qui pêche par défaut : il néglige les variations qui existent d'une commune à l'autre, dans la même strate. Toutefois, il peut très bien se faire que (au moins pour les enquêtes françaises)

ces variations soient beaucoup moins importantes que celles existant entre ménages (ou personnes) de la même commune. C'est même d'autant plus vraisemblable que les strates sont plus nombreuses et homogènes, les communes étant stratifiées suivant la région, la taille, le caractère rural ou la nature de l'industrie dominante. Mais ce sera plus probablement le cas pour telle variable étudiée que pour telle autre.

Ainsi, on peut très bien concevoir que le calcul de V_0 fournisse des indications utiles (sous réserve qu'on n'oublie pas l'existence du terme omis dans $\mathcal{V}\bar{x}$).

6 - EFFET COMBINE DES STRATES ET DES DEGRES DE SONDAGE

Lorsqu'on tire un échantillon très stratifié, -à 1 degré dans les strates constitués dans les grandes villes et à 2 degrés dans les autres strates (comme c'est le cas en France)- , on n'est jamais obligé de se placer dans une position aussi mauvaise que celle du § 3.

Supposons qu'on ait 2 communes par strate dans les strates A B C ... On connait les nombres de ménages par strate, de sorte qu'on calcule une estimation de la forme

$$\bar{x} = \alpha\,\bar{a} + \beta\,\bar{b} + \gamma\,\bar{c} + \ldots$$

où $\alpha\ \beta\ \gamma\ldots$ sont les coefficients connus, et $\bar{a}\ \bar{b}\ \bar{c}\ \ldots$ des moyennes-échantillon (par ménage). Une dichotomie des communes fournit $\bar{a}_1\bar{b}_1\bar{c}_1\ldots$ et $\bar{a}_2\bar{b}_2\bar{c}_2\ \ldots$ séparément, d'où :

$$\bar{x}_1 - \bar{x}_2 = \alpha\,(\bar{a}_1 - \bar{a}_2) + \beta\,(\bar{b}_1 - \bar{b}_2) + \gamma\,(\bar{c}_1 - \bar{c}_2) + \ldots$$

$$\Longrightarrow \quad \mathcal{E}(\bar{x}_1 - \bar{x}_2)^2 = \alpha^2\ \mathcal{E}(\bar{a}_1 - \bar{a}_2)^2 + \beta^2\ \mathcal{E}(\bar{b}_1 - \bar{b}_2) + \gamma^2\ \mathcal{E}(\bar{c}_1 - \bar{c}_2)^2$$

$$+ \ldots$$

si les couples d'écarts $(a_1 - a_2)$ et $(b_1 - b_2)$, etc... sont bien indépendants, c'est-à-dire si la désignation des communes-échantillon a bien été faite dans chaque strate par tirages au sort (avec remise).

On a alors quelque intérêt à estimer la variance en se servant de

$$\omega = \alpha^2\,(\bar{a}_1 - \bar{a}_2)^2 + \beta^2\ (\bar{b}_1 - \bar{b}_2)^2 + \gamma^2\ (\bar{c}_1 - \bar{c}_2)^2 + \ldots$$

au lieu de $(\bar{x}_1 - \bar{x}_2)^2$. En effet :

$\bar{a}_1$ et $\bar{a}_2$ étant 2 valeurs d'une variable de Laplace Gauss d'écart-type σ_a , $\alpha^2\,(\bar{a}_1 - \bar{a}_2)^2$ est $2\,\alpha^2\,\sigma_a^2$ fois un χ^2 à 1 degré de liberté ; ω n'est pas un χ^2 (1), sauf dans le cas particulier où $\alpha\sigma_a = \beta\sigma_b = \gamma\sigma_c = \ldots$ Dans ce cas seulement on peut dire que $\omega(\alpha^2\ \sigma_a^2 + \beta^2\ \sigma_b^2 + \gamma^2\sigma_c^2)^{-1}$ a autant de degrés de liberté qu'il y a de strates.

Pour une "grande région", l'I.N.S.E.E. constitue habituellement une *dizaine de strates* ; nous sommes donc dans le cas d'une estimation raisonnablement bonne de la variance.

---

(1) L'identification entre les fonctions génératrices de moments

$(1 - 2\,t\,\lambda^2)^{-1/2}.\,(1 - 2\,t\,\mu^2)^{-1/2}$ et $(1 - 2\,t\,\rho^2)^{-1}$ est impossible si $\lambda^2 \neq \mu^2$.

Dans le cas général, il paraît raisonnable de penser qu'il en est encore de même, mais cela mériterait réflexion(1).

Une difficulté supplémentaire apparaît à mesure que les $\bar{a}_i$ $\bar{b}_i$ $\bar{c}_i$ ... sont calculés avec des échantillons trop petits pour qu'on ait le droit de les confondre avec des variables de Laplace Gauss ; alors les $(\bar{a}_1 - \bar{a}_2)^2$ et analogues ne sont plus proportionnels à des χ^2 .

Enfin, on doit observer que, bien que $\bar{a}_i$ $\bar{b}_i$ $\bar{c}_i$ ... soient évalués sur l'échantillon, ce qu'on estime au moyen de $\omega/4$ est exactement $\nu\bar{x}$ (on n'isole pas une variance interne et une variance externe). On obtiendrait à peu de choses près $\omega = (\bar{x}_1 - \bar{x}_2)^2$ si on prenait soin de tirer au sort le signe + ou - à associer à $|\bar{a}_1 - \bar{a}_2|$, à $|\bar{b}_1 - \bar{b}_2|$ etc, autrement dit si l'on tirait bien au sort laquelle des 2 communes de chaque strate recevra le numéro 1 (plutôt que le numéro 2) dans le calcul de $\bar{x}_1 - \bar{x}_2$, (et si les strates sont nombreuses).

7 - RESUME

On a indiqué des procédés en vue de se faire (à peu de frais) une idée suffisamment précise de l'ampleur des erreurs d'échantillonnage des sondages courants en France,

---

(1) Il existe une littérature considérable sur la distribution des mélanges de χ^2, mais elle présente trop rarement un caractère pratique.