JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

JACQUES DESABIE

# Sur un problème d'échantillon « optimum »

*Journal de la société statistique de Paris*, tome 97 (1956), p. 130-135

<http://www.numdam.org/item?id=JSFS_1956__97__130_0>

## Sur un problème d'échantillon " optimum " (1)

a) *Rappel : la notion d' « optimum » en théorie des Sondages.*

L'objet de la méthode des Sondages est de fournir, à partir des mesures effectuées sur un échantillon, une estimation des paramètres : moyenne, proportion, caractéristiques de la population étudiée.

Pour estimer un paramètre donné on choisit un « estimateur », c'est-à-dire une fonction des valeurs prises par la (ou les) variable sur les unités échantillon. Par exemple, pour estimer la moyenne de la population, on utilisera l'estimateur :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Pour estimer la variance de la population, on utilisera l'estimateur :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2$$

La valeur prise par l'estimateur pour l'échantillon effectivement désigné par le sort sera retenue comme estimation.

Tout estimateur est une variable aléatoire dont la valeur varie d'un échantillon à l'autre. La loi de probabilité de l'estimateur est d'ailleurs déterminée pour une population et un estimateur donnés par le mode de tirage adopté et l'effectif de l'échantillon; l'estimateur sera d'autant plus satisfaisant que sa distribution sera moins étalée, c'est-à-dire que sa variance sera plus faible.

Le problème réside donc dans le choix simultané d'un mode de tirage et

(1) Cet article a paru dans le premier cahier de l'ADETEM, organisme qui groupe les spécialistes français des études de marchés. Il est reproduit ici avec l'aimable autorisation de cet organisme.

d'un estimateur tels que la variance de l'estimateur soit minimum pour un coût donné. Posé en des termes aussi généraux, le problème n'est pas susceptible d'une solution mathématique simple; il est en revanche susceptible d'une solution lorsque mode de tirage et estimateur sont donnés *a priori* à quelques paramètres près que l'on déterminera en minimisant la variance à coût constant.

b) *Historiquement* le problème a été traité pour la première fois par Neyman : étant donné un univers stratifié (les strates sont donnés *a priori*) et un estimateur linéaire non biaisé comment répartir l'échantillon entre les strates pour obtenir une variance minimum.

Jusqu'à une époque assez récente, le problème du découpage optimum de l'univers en strates n'avait pas fait l'objet d'une étude mathématique. On savait, bien entendu, que les strates devaient être aussi homogènes que possible et des études empiriques avaient montré que si il y avait toujours intérêt à augmenter le nombre de strates, le gain sur la précision était assez rapidement décroissant. A cela se limitaient les indications de la « Science ».

Un problème de détermination des strates optima s'est posé à l'I. N. S. E. E. à l'occasion de la préparation :

— du Recensement agricole;
— d'un échantillon principal (master sample) d'établissements.

Le problème du recensement agricole était le suivant : un questionnaire restreint dit questionnaire général serait rempli par toutes les exploitations. Des questionnaires particuliers, plus complexes, seraient remplis chacun par un échantillon d'exploitations. Il était entendu à l'avance que l'univers des exploitations serait stratifié par départements (ou régions agricoles), que dans chaque département les grosses exploitations seraient toutes incluses dans l'échantillon — un taux de sondage uniforme étant adopté pour les autres exploitations.

Il restait à décider du seuil au-dessus duquel une exploitation serait considérée comme grande dans un département (ou une région agricole) donné.

La variable retenue pour mesurer la taille d'une exploitation était sa surface, et l'on cherchait à rendre maximum la précision de l'estimation de la surface moyenne par exploitations pour *l'ensemble du territoire* (et non par département).

Le problème de l'échantillon principal d'établissements était *identique* à cela près, que les établissements étaient stratifiés par *activité collective* et non par départements, et que l'on avait adopté le nombre de salariés comme mesure de la taille d'un établissement.

La solution mathématique du problème est donnée ci-dessous, la rédaction est appropriée à l'étude de l'échantillon principal d'établissements, ce qui permet d'adopter un langage plus concret, mais ne restreint en rien la généralité du résultat obtenu.

***

Pour une activité collective $A$, on note
$\xi$ la variable additive « nombre de salariés par établissement »
$\xi_0$ la valeur de $\xi$ où a lieu la coupure (à déterminer).

$\varphi(\xi) \cdot d\xi$ la fréquence (relative) élémentaire en $\xi$

$\Phi(\xi) = \int_0^{\xi} \varphi(u) \cdot du$ la fréquence cumulée de 0 jusqu'à $\xi$

$\mu$ le nombre total d'établissements

$\mu_0 = \mu \cdot \Phi(\xi_0)$ le nombre de petits établissements : définis par : $\xi \leqslant \xi_0$

$\mu - \mu_0$ le nombre de grands établissements : définis par : $\xi > \xi_0$

Des $\mu_0$ petits établissements on prélève un échantillon au hasard sans remise d'effectif $m_0 = f\,\mu_0$, $f$ étant la fraction de sondage. La *totalité des* $\mu - \mu_0$ grands établissements est incluse dans l'échantillon.

Le nombre moyen de salariés par établissement de l'activité $A$ est

$$\bar{x} = \int_0^{\infty} \xi \cdot \varphi(\xi) \cdot d\xi \tag{1}$$

et le nombre total des salariés de l'activité $A$.

$$x = \mu \cdot \bar{x} \tag{2}$$

Le nombre moyen de salariés par établissement pour les petits établissements est

$$\bar{x}_0 = \frac{\int_0^{\xi_0} \xi \cdot \varphi(\xi) \cdot d\xi}{\int_0^{\xi_0} \varphi(\xi) \cdot d\xi} = \frac{1}{\Phi(\xi_0)} \int_0^{\xi_0} \xi \cdot \varphi(\xi) \cdot d\xi \tag{3}$$

et le nombre total de salariés des petits établissements est

$$x_0 = \mu_0 \cdot \bar{x}_0 \tag{4}$$

tandis que la variance de $\xi$ entre petits établissements

$$V_0(\xi) = \frac{1}{\Phi(\xi_0)} \int_0^{\xi_0} (\xi - \bar{x}_0)^2 \cdot \varphi(\xi) \cdot d\xi \tag{5}$$

On estimera le nombre total de salariés de l'activité $A$, soit $x$ (formule 2) par :

$$X = \frac{1}{f} \cdot S_1^{m_0} \xi_i + \sum_1^{\mu - \mu_0} \xi_j \tag{6}$$

($S$ désigne une sommation étendue aux $m_0$ unités-échantillon; $\Sigma$ désigne une sommation étendue aux $\mu$-$\mu_0$ unités de la population).

La variance de $X$ est uniquement celle de son premier terme :

$$V(X) = \frac{1}{f^2} \cdot V\left[S_1^{m_0} \xi_i\right] = \frac{m_0}{f^2} V_0(\xi)(1 - f) \tag{7}$$

$$= \frac{\mu(1 - f)}{f} \int_0^{\xi_0} (\xi - \bar{x}_0)^2 \cdot \varphi(\xi) \cdot d\xi$$

Pour une autre activité collective $B$ on aura des notations analogues, la correspondance étant donnée ci-après avec celles définies précédemment pour l'activité $A$ (en général on passe de $A$ à $B$ en prenant la lettre suivante de l'alphabet) :

| Activité A | Activité B | Activité A | Activité B |
|---|---|---|---|
| $\xi$ | $\eta$ | $\bar{x}$ | $\bar{y}$ |
| $\xi_0$ | $\eta_0$ | $\underline{x}$ | $\underline{y}$ |
| $\varphi(\xi)$ | $\chi(\eta)$ | $\bar{x}_0$ | $\bar{y}_0$ |
| $\Phi(\xi)$ | $\psi(\eta)$ | $x_0$ | $y_0$ |
| $\mu$ | $\nu$ | $V_0(\xi)$ | $V_0(\eta)$ |
| $\mu_0$ | $\nu_0$ | | |
| $m_0$ | $n_0$ | X | Y |
| $f$ | $g$ | V (X) | V (Y) |

On suppose que le coût moyen par questionnaire est le même quelle que soit l'importance ou l'activité d'un établissement. Le coût total de l'enquête est donc proportionnel au nombre total d'établissements interrogés $E$.

$$(8) \qquad \begin{aligned} E &= m_0 + (\mu - \mu_0) + n_0 + (\nu - \nu_0) \\ &= \mu\,[1 - (1 - f)\cdot\Phi(\xi_0)] + \nu\,[1 - (1 - g)\cdot\psi(\eta_0)] \end{aligned}$$

*On se propose de déterminer les deux points $\xi_0$ et $\eta_0$ rendant minimum la variance de l'estimation $X + Y$ de $x + y$ pour un coût total donné, c'est-à-dire pour un nombre total donné de questionnaires dans l'ensemble des deux activités $A$ et $B$.*

Ces deux valeurs $\xi_0$ et $\eta_0$ sont telles que le système en $d\xi_0$, $d\eta_0$

$$(9) \quad d[V(X) + V(Y)] = \frac{\delta V(X)}{\delta \xi_0} d\xi_0 + \frac{\delta V(Y)}{\delta \eta_0} \cdot d\eta_0 = 0$$

$$dE = \mu \frac{\delta}{\delta \xi_0}[1 - (1 - f)\,\Phi(\xi_0)] \cdot d\xi_0 + \nu \frac{\delta}{\delta \eta_0}[1 - (1 - g)\,\psi(\eta_0)] \cdot d\eta_0$$

soit compatible, ce qui n'a lieu que si les coefficients des différentielles sont proportionnels.

On a :

$$(10) \quad \frac{\delta V(X)}{\delta \xi_0} = \frac{\mu(1-f)}{f}(\xi_0 - x_0)^2 \cdot \varphi(\xi_0) - 2\frac{\mu(1-f)}{f}\int_0^{\xi_0}(\xi - x_0)\frac{\delta \bar{x}_0}{\delta \xi_0}\cdot\varphi(\xi)\cdot d\xi$$

le dernier terme étant nul, et

$$(11) \qquad \frac{\delta}{\delta \xi_0}[1 - (1 - f)\cdot\Phi(\xi_0)] = (1 - f)\cdot\varphi(\xi_0)$$

On a deux expressions analogues pour les termes faisant intervenir les quantités relatives à l'activité $B$, d'où la condition :

$$(12) \qquad \frac{1}{f}(\xi_0 - \bar{x}_0)^2 = \frac{1}{g}(\eta_0 - \bar{y}_0)^2$$

qui se réduit, dans le cas où les deux fractions de sondage $f$ et $g$ sont égales à :

$$(13) \qquad \boxed{\xi_0 - \bar{x}_0 = \eta_0 - \bar{y}_0}$$

Cette condition se généralise immédiatement à un nombre quelconque d'activités collectives (1).

*Détermination pratique des* $\xi_0$, $\eta_0$...

La variable $\xi$ varie évidemment, non pas d'une manière continue comme il a été supposé ci-dessus, mais par unité pour les premières valeurs, par plusieurs unités ensuite. On ne fixera pratiquement la coupure qu'en une limite de classe.

Le tableau à établir pour chaque activité est le suivant :

| Limite supérieure de classe | Nombre cumulé d'établissements | Nombre cumulé de salariés | Nombre moyen de salariés | Écarts |
|---|---|---|---|---|
| $\alpha_1$ | $\mu_1$ | $x_1$ | $\bar{x}_1 = x_1/\mu_1$ | $\alpha_1 - x_1$ |
| $\alpha^2$ | $\mu_2$ | $x_2$ | $\bar{x}_2 = x_2/\mu_2$ | $\alpha_2 - x_2$ |
| $\alpha_3$ | $\mu_3$ | $x_3$ | $\bar{x}_3 = x_3/\mu_3$ | $\alpha_3 - x_3$ |
| ..... | ..... | ..... | ..... | ..... |
| $\alpha_k$ | $\mu_k = \mu$ | $x_k = x$ | $\bar{x}_k = \bar{x} = x/\mu$ | $\alpha_k - x_k$ |

S'étant donné la valeur de l'écart pour une activité, on calculera le nombre d'établissements à soumettre à l'enquête dans chaque activité et pour l'ensemble. On pourra trouver après quelques essais seulement la valeur convenable de l'écart telle que le nombre total d'établissements à visiter soit le nombre désiré.

La solution peut d'ailleurs être obtenue à l'aide d'une construction géométrique assez simple mise au point par M. Duprat.

*REMARQUE :* Le tableau ci-dessus suppose connu le nombre de salariés de chaque établissement, supposition réalisée en fait. Il ne s'agira pas d'estimer le nombre connu $x$, par exemple : mais le chiffre d'affaires, la masse des salaires versés par les établissements.

Toutefois le chiffre d'affaires, les salaires versés par un établissement sont en étroite corrélation avec le nombre de salariés qu'il emploie. Le plan de Sondage optimum pour l'évaluation du nombre de salariés sera donc satisfaisant pour l'estimation du chiffre d'affaires ou de la masse des Salaires.

### GÉNÉRALISATIONS

*a*) On a déjà vu que le résultat se généralisait immédiatement au cas où le taux de sondage pour les petits établissements était variable suivant le groupe d'activité tout en étant donné *a priori*. On n'a pas traité en revanche le cas où ces taux de sondages seraient eux-mêmes des paramètres variables à déterminer. On n'a pas davantage étudié ce qui se passerait si l'on abandonnait l'hypothèse d'un taux de sondage à 100 % pour les gros établissements.

*b*) Le résultat se généralise immédiatement au cas où le coût de l'enquête serait différent suivant les activités. On n'a pas étudié le cas où ce coût serait variable avec la taille de l'établissement.

(1) Il est à noter que lorsque la relation $\xi_0 - \bar{x}_0 = \text{cte...}$ est vérifiée pour l'ensemble des groupes d'activité, elle est également vérifiée pour tout sous-ensemble. Le mode de découpage est donc optimum pour une enquête sur l'industrie seule, ou le commerce seul.

*c*) La généralisation serait sans doute immédiate si l'on étudiait une variable en corrélation linéaire à l'intérieur de chaque groupe d'activités, avec la « taille » de l'unité.

### APPLICATION NUMÉRIQUE

Une application numérique a été entreprise en ce qui concerne les exploitations agricoles et les établissements.

On s'est aperçu alors que la Condition $\xi_o - \bar{x}_o = \text{cte}$ était très peu différente de la Condition $\xi_o = \text{cte}$.

En effet, la Superficie moyenne des exploitations de moins de 50 hectares par exemple varie peu d'un département à l'autre (et même d'une région agricole à l'autre) *de sorte que la Condition est* à peu près satisfaisante en prenant une même superficie comme coupure *pour tous les départements* — il en est de même d'ailleurs en ce qui concerne les établissements.

La conclusion pratique est donc qu'il faudra prendre comme coupure la même surface dans tous les départements — (ou le même nombre de salariés pour tous les groupes d'activité).

Nous rencontrons ici un exemple remarquable où la Théorie conduit à un résultat parfaitement simple et utilisable encore qu'assez peu « joli ».

| DÉPARTEMENTS | NOMBRES CUMULÉS | SURFACES CUMULÉES | SURFACE MOYENNE | COUPURE surface moyenne |
|---|---|---|---|---|
| AIN : | | | | |
| Moins de 1 ha | 2.555 | 1.283 | 0,50 | 0,50 |
| — 2 ha | 5.030 | 4.796 | 0,95 | 1,05 |
| — 3 ha | 7.624 | 10.975 | 1,44 | 1,56 |
| — 5 ha | 12.881 | 31.291 | 2,43 | 2,57 |
| — 10 ha | 23.380 | 105.507 | 4,51 | 5,49 |
| — 20 ha | 30.522 | 201.368 | 6,60 | 13,40 |
| — 40 ha | 32.991 | 266.842 | 8,09 | 31,91 |
| — 50 ha | 33.398 | 284.606 | 8,52 | 41,48 |
| — 100 ha | 33.938 | 319.468 | 9,41 | 90,59 |
| — 200 ha | 34.046 | 334.014 | 9,81 | 190,19 |
| — 500 ha | 34.109 | 350.482 | 10,28 | 489,72 |
| TOTAL | 34.119 | 362.710 | 10-63 | |
| AISNE : | | | | |
| Moins de 1 ha | 538 | 272 | 0,51 | 0,49 |
| — 2 ha | 1.141 | 1.107 | 0,97 | 1,03 |
| — 3 ha | 1.726 | 2.500 | 1.45 | 1,55 |
| — 5 ha | 2.742 | 6.374 | 2,32 | 2,68 |
| — 10 ha | 4.513 | 18.910 | 4,19 | 5,81 |
| — 20 ha | 7.356 | 59.740 | 8,12 | 11,88 |
| — 40 ha | 13.061 | 143.642 | 13,86 | 26,14 |
| — 50 ha | 11.101 | 176.243 | 15,88 | 34,12 |
| — 100 ha | 12.395 | 264.165 | 21,31 | 78,69 |
| — 200 ha | 13.124 | 364.804 | 27,80 | 172,20 |
| — 500 ha | 13.552 | 486.204 | 35,88 | 464,12 |
| TOTAL | 13.582 | 509.160 | 37,49 | |
| BASSES-ALPES : | | | | |
| Moins de 1 ha | 459 | 210 | 0,46 | 0,54 |
| — 2 ha | 922 | 857 | 0,93 | 1,07 |
| — 3 ha | 1.323 | 1.804 | 1,36 | 1,64 |
| — 5 ha | 2.084 | 4.752 | 2,28 | 2,72 |
| — 10 ha | 3.749 | 16.717 | 4,46 | 5,54 |
| — 20 ha | 5.941 | 47.970 | 8,07 | 11,93 |
| — 40 ha | 7.645 | 95.309 | 12,47 | 27,53 |
| — 50 ha | 8.053 | 113.169 | 14,05 | 35,95 |
| — 100 ha | 8.988 | 176.865 | 19,68 | 80,32 |
| — 200 ha | 9.415 | 234.610 | 24,92 | 175,08 |
| — 500 ha | 9.511 | 261.776 | 27,52 | 472,48 |
| TOTAL | 9.535 | 293.608 | 30-79 | |

M. DESABIE.