Échantillonnage à probabilités inégales (plan PPS)

Pierre Duchesne

November 28, 2018

Introduction

- Dans un plan à probabilités inégales, toutes les probabilités d'inclusion d'ordre un, les π_k, $k \in U$, sont potentiellement différentes.
- En fait ce n'est pas complètement nouveau. On a vu que pour l'échantillonnage dans le plan STSI, on devait parfois utiliser des probabilités d'inclusion qui ne sont pas égales pour toutes les unités $k \in U$.
- On rappelle par exemple la règle de Neyman: $n_h \propto N_h S_{yU_h}$, c'est-à-dire:

$$\frac{n_h}{N_h} \propto S_{yU_h}.$$

Implications d'avoir des π_k différents

- Ainsi, dans le plan STSI, pour une unité $k \in U_h$:

$$f_h = \frac{n_h}{N_h} = \pi_k.$$

- Si la variance dans la strate est élevée, π_k sera aussi élevée. Un grand S_{yU_h} est associée à de l'incertitude, il y aura plus de chances de choisir les unités dans de telles strates.
- Inversement, si la variance dans la strate est faible, les π_k seront aussi faibles.

Plan πPS

- La conséquence ultime de cette idée: attribuer directement, lors de la conception du plan de sondage, à chaque unité k sa propre probabilité d'inclusion π_k.
- Deux grandes catégories de plans de sondage dans cet esprit.
- Plan πPS: les probabilités d'inclusion sont fournies, et on cherche un plan *sans remise* qui respecte les π_k.
- Les plans πPS sont généralement compliqués à étudier.

Plan PPS

- Les plans PPS sont les versions *avec remise* des plans πPS.
- PPS: *Probability Proportional to Size*, autrement dit, les probabilités seront proportionnelles à une variable de taille.
- Comme le nom l'indique, pour une variable d'intérêt y, il faudra une variable auxiliaire x (variable de taille).
- Habituellement, comme on le verra, si une bonne corrélation existe entre y et x, alors un plan PPS mènera à des estimations avec faibles variances.

Mise en oeuvre d'un plan PPS

- Considérons la variable de taille x_k, $k \in U$.
- On procède de la manière suivante:

$$p_k = \frac{x_k}{\sum_U x_k} = \frac{x_k}{N\bar{x}_U}, \ k = 1, 2, \ldots, N.$$

- On note que l'on a $\sum_U p_k = 1$.
- On effectue m tirages indépendamment.
- On tirage i, on sélectionne l'unité k avec probabilité p_k. On fait cela pour les m tirages, $i = 1, \ldots, m$.
- On insiste que les tirages sont sans remise et indépendants.

Échantillon

- Soient $k_1, k_2, \ldots, k_i, \ldots, k_m$ les unités choisies.
- On observe y_{k_i}, $i = 1, \ldots, m$.
- Le nombre d'unités distinctes est au plus m: en effet, il peut y avoir des répétitions.
- L'estimateur proposé est le suivant:

$$\hat{t}_{pr} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{p_{k_i}} = \frac{N\bar{x}_U}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{x_{k_i}}.$$

- Pourquoi cet estimateur? L'idée est que si $y_k \approx Cx_k$, $k \in U$, alors l'estimateur sera stable.

Motivation pour l'estimateur

- On procède comme on a déjà fait. Considérons le cas limite où l'on a $y_k \equiv Cx_k, \forall k \in U$.
- On aura alors:

$$\begin{aligned}
\hat{t}_{pr} &= \frac{N\bar{x}_U}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{x_{k_i}}, \\
&= \frac{N\bar{x}_U}{m}\sum_{i=1}^{m}\frac{Cx_{k_i}}{x_{k_i}}, \\
&= \frac{N\bar{x}_U}{m}\sum_{i=1}^{m} C, \\
&= \frac{\sum_U x_k}{m} mC, \\
&= \sum_U Cx_k = \sum_U y_k = t_y.
\end{aligned}$$

Motivation pour l'estimateur (suite)

- Ainsi, si $y_k \equiv Cx_k, \forall k \in U$, on aurait que:

$$\hat{t}_{pr} = \sum_U y_k = t_y.$$

- L'estimation serait alors parfaite, quelque soit $k_1, k_2, \ldots, k_i, \ldots, k_m$.
- La variance serait alors nulle.
- En pratique la relation entre y_k et x_k ne sera évidemment pas parfaite.

L'estimateur $\hat{t}_{pr}$ n'est pas l'estimateur Horvitz-Thompson

- On note que p_k n'est pas la probabilité d'inclusion car suite à m tirages il pourrait y avoir des répétitions.
- On peut calculer π_k:

$$P(s \ni k) = 1 - P(s \text{ ne contient pas } k) = 1 - (1 - p_k)^m$$

- Posons $f(x) = 1 - (1 - x)^m$. On a que $f(0) = 0$, et $f'(x) = m(1 - x)^{m-1}$, et pour x proche de 0:

$$f(x) \approx f(0) + f'(0)x = mx$$

- Si p_k est petit on a alors que $\pi_k \approx mp_k$.
- Alternativement, on pourrait considérer $\sum_s y_k/\pi_k$ avec $\pi_k = 1 - (1 - p_k)^m$.

Réalisation d'un plan PPS

- Supposons que l'on dispose de $x_1, x_2, \ldots, x_k, \ldots, x_N$, avec $x_k > 0, \forall k \in U$.
- On sait que si $U \sim U[0,1]$, si $[a,b] \subset [0,1]$, alors $P(U \in [a,b]) = b - a$.
- Si suffit de former les sommes partielles: $S_1 = x_1$, $S_2 = x_1 + x_2$, et plus généralement:

$$S_k = \sum_{i=1}^{k} x_i.$$

- On considère les intervalles $(0, S_1/t_x], (S_1/t_x, S_2/t_x], \ldots, (S_{k-1}/t_x, S_k/t_x], \ldots, (S_{N-1}/t_x, 1]$.
- On génère $U_1, U_2, \ldots, U_m$ des variables aléatoires uniformes sur $[0,1]$.
- Au tirage i, si $u_i \in (S_{k-1}/t_x, S_k/t_x]$, on choisit l'unité k.
- Au tirage i, la probabilité de sélectionner l'unité k est bien x_k/t_x.