Proposition d'exercice de sondage

J. Collet [a] [b]

[a] *EDF R&D*
1, avenue du Général de Gaulle
F-92141 Clamart CEDEX
Jerome.Collet@edf.fr
[b] *École Polytechnique Universitaire de Lille*
Avenue Paul Langevin
59655 Villeneuve d'Ascq CEDEX

RÉSUMÉ. *Cet article décrit un TP permettant de mettre en oeuvre une bonne partie des méthodes usuelles de sondages, tant du point de vue de l'échantillonnage que de celui de l'estimation*

ABSTRACT. *This paper describes an exercise, for which one haves to use many of the usual polling methods*

MOTS-CLÉS : *sondages, enseignements, exercices*

KEYWORDS: *polls, teaching, exercises*

Si, lors d'un cours sur les méthodes de sondage, par exemple sur les notions exposées dans [ARD 94, TIL 01], on veut insister sur leur mise en pratique, une difficulté se présente : la construction d'exercices. Nous proposons ici un exercice léger à organiser, durant environ 3 heures, et qui fait appliquer un assez grand nombre des méthodes usuelles.
L'exercice consiste à estimer le nombre de mots d'un livre donné.
Cet exercice a été proposé à l'École Polytechnique Universitaire de Lille, en troisième année. Le cours (6 séances de 4 heures, réparties sur 3 semaines) était ouvert à toutes les options de l'école, les élèves ayant choisi ce cours venaient d'options aussi diverses que « Génie Informatique et Statistique » et « Industries AgroAlimentaires ». L'exercice semble avoir été efficace.

1. L'exercice effectivement proposé

Le livre proposé est « Le tour du monde en 80 jours », de Jules Verne. L'intérêt de ce roman assez ancien est qu'il est dans le domaine public, disponible sur Internet, ce qui permet de compter les mots automatiquement, et d'avoir une référence. L'édition proposée est l'édition Folio Junior, qui comprend quelques illustrations (couvrant

chacune une page).
La première étape consiste à ,préciser la population. La définition retenue était : tous les mots du teste, sans la table des matières (de la page 9 à la page 294).
Ensuite, il apparaît qu'il est plus efficace de choisir d'abord des pages, puis des lignes dans chaque page choisie. Il faut donc compter toutes les lignes des pages choisies, mais pas des autres : l'avantage du tirage à plusieurs degrés est alors évident.
Il est évident qu'il y a 3 types de pages : les pages de saut de chapitre (car dans ce livre presque tous les sauts de chapitre sont sur une page), les pages avec image, et les pages « normales ». La table des matières nous donne la liste des pages de saut de chapitre, ce qui permet d'utiliser la méthode de stratification. En revanche, comme ce livre ne comprend pas de table des illustrations, le traitement des pages avec images montrera l'utilité de la post-stratification.
Il apparaîtra aussi utile de distinguer les lignes complètes des incomplètes. Cependant, cette distinction ne sera faite que sur les pages normales, qui sont à la fois les plus nombreuses et les plus importantes.
Tous les tirages sont des tirages systématiques.

2. Pages normales

2.1. *Échantillonnage : difficulté pratique*

Dans le cas d'un sondage à plusieurs degrés, il vaut mieux avoir un plus fort taux de sondage au premier degré qu'au second. De plus, dans le cas d'un roman, la présence d'un dialogue risque d'affecter toutes les lignes d'une page. Nous prendrons donc 4 pages, et 2ou 4 lignes par page (selon que l'on sépare les lignes complètes des incomplètes).
L'échantillonnage des pages pose des problèmes de décalage. Il faut d'abord soustraire du nombre de pages le nombre de sauts de chapitre, **supposé** égal au nombre de chapitres. Le tirage systématique donne un numéro parmi les pages normales. Pour connaître le vrai numéro de la page tirée, il faut tenir compte du nombre de pages sautées depuis le début du livre, ajouter ce nombre de pages, et corriger une seconde fois si nécessaire.

2.2. *Eswtimation du nombre de mots par ligne : intérêt d'une seconde stratification*

On constate, par exemple en comparant avec la **vraie** valeur, que ne pas séparer les lignes complètes des incomplètes donne des estimations **très** imprécises. Or, il est assez facile de compter toutes les lignes complètes d'une page donnée pour en choisir 2, et faire de même pour les incomplètes.

3. Pages de changement de chapitre

Cette population est peu importante : il suffit de prendre 2 pages, 2 lignes par page, et la suite du calcul ne pose pas de problème du type de celui rencontré pour les pages normales.

4. Pages avec image : intérêt de la post stratification

Les pages avec image, dont nous n'avons pas de table, montrent l'intérêt de la post-stratification.
De plus, elles permettent de constater qu'il serait parfois nécessaire de faire des développements théoriques spécifiques. En effet, il pourrait être intéressant d'**estimer** le nombre de pages avec images. cela serait possible à l'aide d'un sondage sur la population des pages. Cependant, le calcul de la variance de l'estimateur post-stratifié ferait alors intervenir la variance de l'estimation du nombre de pages avec image. C'est évidemment possible, mais nécessiterait des développements spécifiques.
C'est pourquoi il vaut mieux renoncer à cette estimation, pour lui préférer un comptage exhaustif qui prend à peu près 2 minutes.

5. Quelques remarques pratiques

– Avoir tiré au préalable des nombres au hasard (suivant une uniforme sur [0,1]) permet de connaître le résultat du TP à l'avance.

– Il vaut mieux prévoir plusieurs exemplaires du livre, par exemple 1 pour 6.

– Il faut absolument éviter un livre divisé en parties et chapitres. En effet, dans ce cas, la recherche de la $n^{\text{ième}}$ page normale sera très fastidieux. Une autre solution serait alors alors de changer de méthode de tirage..

6. Bibliographie

[ARD 94] ARDILLY P., *Les techniques de sondage,* Éditions Technip, 1994.

[TIL 01] TILLÉ Y., *Théorie des sondages,* Éditions Dunod, 2001.