REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

Application des nombres de Stirling de 2e espèce à un problème de sondage

Revue de statistique appliquée, tome 15, n° 4 (1967), p. 35-46

<http://www.numdam.org/item?id=RSA_1967__15_4_35_0>

APPLICATION DES NOMBRES DE STIRLING DE 2e ESPÈCE A UN PROBLÈME DE SONDAGE

P. THIONET

Il n'est guère de statisticien qui ne soit familiarisé avec les tirages de boules dans une urne, tantôt avec remise (tirages de Bernoulli), tantôt sans remise (tirages dits exhaustifs car ils tendent à épuiser le contenu de l'urne). En statistique élémentaire, les boules sont de plusieurs couleurs ; et l'on cherche à estimer la proportion de boules (disons) noires sans vider l'urne.

Avec la théorie élémentaire des sondages, on passe au cas d'une variable X réelle à valeurs sur chaque boule de l'urne ; en particulier, si la variable ne peut prendre que les valeurs 0 ou 1, on retrouve le premier problème. Finalement il s'agit d'estimer la moyenne de X à l'aide des valeurs de X sur les boules échantillon.

Les tirages avec remise aboutissant fatalement à tirer plusieurs fois la même boule, il est bien naturel d'en venir à calculer la moyenne de X sur les seules boules distinctes tirées. Ce n'est certes pas aussi bon qu'un tirage exhaustif, où on est sur de n'avoir dans l'échantillon que des boules distinctes ; mais on peut s'attendre à avoir ainsi un estimateur meilleur que si l'on fait la moyenne de X sur toutes les boules tirées, distinctes ou non.

La formulation précise d'un tel problème, l'obtention des résultats corrects, tout ceci s'est fait dans un certain désordre et a été publié de façon dispersée, voici dix ans, en Inde et par l'auteur (INSEE, Etude Théorique n° 7). Le présent article donne une présentation très élaborée de cette petite théorie, et d'abord un chapitre d'Analyse Combinatoire peu connu sur quoi elle repose. On démontre notamment la formule de la variance, donnée sans preuves par RAO (et qui repose sur une formule combinatoire).

I - RAPPEL : NOMBRE DE STIRLING(1)

On appelle nombres de Stirling de première espèce s(n , k), les coefficients des puissances t^k du développement de

$$(t)_n = t(t-1)(t-2)\ldots(t-n+1).$$

Par exemple :

$$(t)_1 = t \; ; \; (t)_2 = -t + t^2 \; ; \; (t)_3 = 2t - 3t^2 + t^3$$

fournissent les lignes n = 2 et n = 3 du tableau ci-après des s(n , k)

Matrice s :

n =						
1	1					
2	-1	1				
3	2	-3	1			
4	-6	11	-6	1		
5	24	-50	35	-10	1	
6	-120	279	-225	85	-15	
7	.....					

On peut inverser la matrice s ci-dessus (arrêtée à nxn, ∀n et complétée de zéros). La matrice $S = s^{-1}$ a pour termes S(n , j) les nombres de Stirling de 2e espèce. Ce sont les coefficients de $(t)_j$ dans t^n exprimé linéairement en fonction des $(t)_1(t)_2 \ldots (t)_j \ldots (t)_n$, par exemple

$$t^2 = (t)_1 + (t)_2 \; ; \; t^3 = (t)_1 + 3(t)_2 + (t)_3 \quad \text{etc.}$$

Matrice S :

n = \ j =	1	2	3	4	5	6	7 ...
1	1						
2	1	1					
3	1	3	1				
4	1	7	6	1			
5	1	15	25	10	1		
6	1	31	90	65	15	1	
7	.....						

Références : Deux livres récents d'Analyse Combinatoire les étudient :

DAVID (F.N.)- BARTON (D.E.) : Combinatorial Chance (1962 - p. 294-5)

RIORDAN (J.) - An Introduction to Combinatorial Analysis (1958- p. 32-42 à 45 - 75 - 91 à 99).

Par contre on ne trouve aucune mention de ces nombres dans les livres de Calcul des Probabilités (même FELLER).

---

(1) STIRLING (James) Methodus Differentialis (1730) p. 6-10. On évitera de confondre ces nombres avec la Formule de Stirling donnant une valeur approchée de n !

2 - RELATIONS DE RECURRENCE - AUTRE NOTATION COURANTE

On vérifie facilement les relations de récurrence suivantes :

$s(n+1\ ,\ j) = s(n\ ,\ j-1) - n\,s(n\ ,\ j)$

$S(n+1\ ,\ j) = S(n\ ,\ j-1) + j\,S(n\ ,\ j)$ qu'on utilisera plus loin.

Les nombres de 2e espèce, qui ont diverses applications dans les distributions de probabilité, sont souvent désignés par une notation de la théorie des différences :

$$\frac{\Delta^j 0^n}{j!} = S(n\ ,\ j)$$

Ceci signifie que la formule générale de Newton :

$$f(x) = \sum_j (x)_j \frac{\Delta^j f(0)}{j!}$$

a été particularisée à $f(x) = x^n$; donc $f(0) = 0^n$.

On trouve (dans les mêmes livres) la définition des nombres dits de Bernouilli généralisés (extension des coefficients du binome), classe à laquelle appartiennent les nombres de Stirling.

3 - UNE DISTRIBUTION DE PROBABILITE DE LA THEORIE DES SONDAGES :

Reprenons la fonction génératrice des $S(n\ ,\ k)$

$$t^n = \sum_{j=1}^{n} S(n\ ,\ j)\,(t)_j$$

Faisons $t = N$ et divisons par N^n les deux membres, il vient

$$1 = \sum_{j=1}^{n} S(n\ ,\ j)\,\frac{N(N-1)\ \ldots\ (N-j+1)}{N^n}$$

ou encore

$$1 = \sum_{j=1}^{n} S(n\ ,\ j)\,\frac{C_n^j j!}{N^n}$$

De la forme $\sum_{j=1}^{n} p_j = 1$, avec $p_j = S(n\ ,\ j)\,C_n^j j!/N^n$

$P_j = \text{Prob}(X = j)$ $\quad j = 1,\ 2,\ 3,\ \ldots\ n$

Cette variable aléatoire X se rencontre notamment dans un problème d'occupation : étant donné N boites et n jetons, répartis au hasard entre les boites, X est le nombre de boites non vides.

En transposant ce problème en un problème (isomorphe) de n tirages (avec remise) de boules dans l'urne à N boules, X est le nombre de boules distinctes tirées en n coups : $X = n_d$.

4 - PROBLEME INVERSE

On sait que le problème des tirages de Bernoulli dans une urne possède un autre aspect (en quelques sorte inverse) : celui de la loi de Pascal : Au lieu d'opérer n tirages donnant un nombre aléatoire B de boules blanches, on se fixe le nombre b de boules blanches à tirer ; et c'est le nombre de tirages n = Y qui devient aléatoire ; on dit aussi qu'on a un phénomène d'attente.

Pareillement, on peut supposer que le nombre de boules distinctes à tirer est fixé d'avance, soit r ; alors le nombre de tirages nécessaire pour attendre r (avec la même urne) est aléatoire, soit $n = S_r$; c'est un autre phénomène d'attente. Il est étudié dans FELLER (T. 1)[1] : page 210. § 3 d : Waiting time in sampling (la suite en exercice, p. 224. Ex. 24-25). Alors il n'est plus question de nombres de Stirling.

FELLER trouve que S_r a une espérance mathématique de l'ordre de $N\mathrm{Log}(N/N - r + 1)$, exactement :

$$\mathcal{E}(S_r) = N\left[\frac{1}{N} + \frac{1}{N-1} + \frac{1}{N-2} + \ldots + \frac{1}{N-r+1}\right]$$

$$\mathcal{V}(S_r) = N\left[\frac{1}{(N-1)^2} + \frac{2}{(N-2)^2} + \ldots + \frac{r-1}{(N-r+1)^2}\right]$$

5 - PROBLEMES DE SONDAGE

On sait que la variance $\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ du sondage par tirage exhaustif (toutes unités tirées distinctes) est inférieure à la variance du sondage bernoullien : σ^2/n.

Supposons que le procédé d'échantillonnage soit bernoullien ; on peut songer à améliorer la variance :

- soit en prolongeant les tirages jusqu'à obtention de n unités distinctes (d'où S_n tirages)

- soit en acceptant le résultat du tirage bernoullien (de taille n), mais en modifiant l'estimateur. Soit $\overline{X} = \sum_{(n)} x_i/N$ la vraie moyenne ; au lieu de l'estimateur

$$\overline{x} = \sum_{(n)} x_i/n$$

on peut user de

$$\overline{x} = \sum_{d} x_i/n_d$$

n_d étant le nombre d'unités distinctes tirées et $\sum_d x_i$ la somme de leurs x . Dans ce qui suit, nous étudierons $\overline{x}'$.

Remarques

1/ Il arrive que la prolongation des tirages jusqu'à S_n soit matériellement impraticable, ou accroisse notablement le coût du sondage.

---

(1) FELLER (W) - An Introduction to Probability Theory and its Applications (3e éd. (1957).

2/ Il est possible d'envisager d'autres estimateurs que $\bar{x}'$, de la forme notamment

$$f(n_d)\,\bar{x}' / \mathcal{E} f(n_d) = \bar{x}''$$

Il est facile de voir que

$$\mathcal{E}(\bar{x} \mid n_d) = \bar{X} \quad ; \quad \forall n_d \; ;$$

donc $$\mathcal{E}(\bar{x}') = \bar{X} \; ;$$

par suite

$$\mathcal{E}(\bar{x}'' \mid n_d) = \frac{(n_d)}{\mathcal{E}(n_d)} \bar{X} \; , \quad \forall n_d \; ;$$

donc $$\mathcal{E}(\bar{x}'') = \bar{X}$$

Tous ces estimateurs sont sans biais ; et il serait possible de choisir parmi eux celui qui rend minimum la variance, liée par quelque condition de coût donné ; mais ceci nous écarte du but de cette étude, c'est-à-dire l'emploi des nombres de Stirling.

6 - PROPRIETES DE L'ESTIMATEUR $\bar{x}'$

A) Absence de biais : Comme on vient de le dire, $\mathcal{E}(\bar{x}' \mid n_d) = \bar{X}$; donc

$$\mathcal{E}(\bar{x}' \; , \; \forall n_d) = \bar{X}$$

B) Variance de $\bar{x}'$: On a manifestement

$$\mathcal{V}(\bar{x}' \mid n_d) = \frac{\sigma^2}{n_d} \frac{N - n_d}{N - 1} = \frac{N\sigma^2}{N-1} \left(\frac{1}{n_d} - \frac{1}{N} \right)$$

donc

$$\mathcal{V}(\bar{x}' \mid \forall n_d) = \sum_1^n \mathcal{V}(\bar{x}/n_d) \,.\, \text{prob}(n_d)$$

$$= \frac{N\sigma^2}{N-1} \mathcal{E}(n_d^{-1}) - \frac{\sigma^2}{N-1}$$

comme $$n_d \leqslant n$$

$$\frac{1}{n_d} \geqslant \frac{1}{n}$$

on a

$$\mathcal{V}(\bar{x} \mid n_d) \geq \frac{\sigma^2}{n} \frac{N - n}{N - 1} \; , \quad \text{donc} \quad \mathcal{V}(\bar{x}') \geq \frac{\sigma^2}{n} \frac{N - n}{N - 1}$$

donc la variance de $\bar{x}'$ est toujours supérieure à celle du sondage exhaustif. Expression exacte de $\mathcal{V}(\bar{x}')$: Pour obtenir $\mathcal{V}(\bar{x}')$, il faut connaître le moment d'ordre (-1) de la distribution de n_d définie ci-dessus :

$$\text{Prob}(n_d = j) = p_j = S(n\ ,\ j)C_N^j j!/N^n$$

7 - THEOREME

$$N^n \mathcal{E}(n_d^{-1}) = N^{n-1} + (N-1)^{n-1} + (N-2)^{n-1} + \ldots + 2^{n-1} + 1$$

En effet :

$$N^n\, \mathcal{E}(n_d^{-1}) = \sum_{j=1}^{n} S(n\ ,\ j)\ C_N^j (j-1)!$$

Mais: (formule de récurrence)

$$S(n\ ,\ j) = S(n-1\ ,\ j-1) + jS(n-1\ ,\ j) = jS(n-1\ ,\ j) + S(n-1\ ,\ j-1)$$

avec

$$S(n\ ,\ n) = S(n-1\ ,\ n-1)$$

$$S(n\ ,\ 1) = \qquad\qquad S(n-1\ ,\ 1)$$

d'où

$$N^n\, \mathcal{E}(n_j^{-1}) = \sum_{j=1}^{n-1} S(n-1\ ,\ j)\ C_N^j\, j! + \sum_{j=2}^{n} S(n-1\ ,\ j-1)\ (j-1)!\ C_N^j$$

Portons au 2ème membre :

$$C_N^j = C_{N-1}^{j-1} + C_{N-2}^{j-1} + \ldots$$

et

$$\sum_{j=1}^{n-1} S(n-1\ ,\ j)\ C_N^j j! = N^{n-1}\ ;$$

$$\sum_{j=1=1}^{n-1} S(n-1\ ,\ j-1)!\ C_{N-1}^{j-1} = (N-1)^{n-1}\ ;\quad \text{etc.}$$

Il vient

$$\boxed{N^n\, \mathcal{E}(n_d^{-1}) = N^{n-1} + (N-1)^{n-1} + (N-2)^{n-1} + \ldots}$$

Reste à voir à quel terme s'arrête le développement. Pour N = 4, n = 3, on a

$$S(3,1) = 1\ ;\ S(3,2) = 3\ ;\ S(3,1) = 1.$$

$$1 = \frac{S(3,1)}{4^3} C_4^1 1 + \frac{S(3,2)}{4^3}\ C_4^2 2! + \frac{S(3,3)}{4^3}\ C_4^3 3! = \frac{4}{64} + \frac{36}{64} + \frac{24}{64}$$

Donc

$$4^3 \mathcal{E}(n_d^{-1}) = 4.1 + 36.\ \frac{1}{2} + 24.\ \frac{1}{3} = 4 + 18 + 8 = 30$$

$$N^{n-1} + (N-1)^{n-1} + (N-2)^{n-1} + \ldots = 4^2 + 3^2 + 2^2 + 1^2 = 16 + 9 + 4 + 1 = 30$$

c'est-à-dire que le développement ne s'arrête qu'à 1^{n-1}.

Détail du calcul :

$$4^3\mathcal{E}(n_d^{-1}) = S(3,1)C_4^1\frac{1}{1} + S(3,2)C_4^2\frac{2!}{2} + S(3,3)C_4^3\frac{3!}{3}$$

$$= S(2,1)C_4^1 + S(2,2)C_4^2 2$$
$$+ S(2,1)C_4^2.1 + S(2,2)C_4^3 2!$$
$$= S(2,1)C_4^1 + S(2,2)C_4^2.2$$
$$+ S(2,1)\,(C_3^1 + C_2^1 + C_1^1) + S(2,2)\,(C_3^2 + C_2^2)\,2!$$
$$= S(2,1)C_4^1 + S(2,2)C_4^2 2 \qquad \text{ou} \quad 4^2$$
$$+ S(2,1)C_3^1 + S(2,2)C_3^2 2 \qquad 3^2$$
$$+ S(2,1)C_2^1 + S(2,2)C_2^2 2 \qquad 2^2$$
$$+ S(2,1)C_1^1 \qquad 1^2$$

Autre exemple : N = 7, N = 5

$$S(5,j) = (1,\ 15,\ 25,\ 10,\ 1)$$

$$7^5\mathcal{E}(n^{-1}) = S(5,1)C_7^1 = S(5,2)C_7^2 + S(5,3)C_7^3 2! + S(5,4)C_7^4 3! + S(5,5)C_7^5 4!$$

On peut en tirer :

$$7^5\mathcal{E}(n_d^{-1}) = S(4,1)C_7^1 + S(4,2)C_7^2 2! + S(4,3)C_7^3 3! + S(4,4)C_7^4 4!$$
$$+ S(4,1)C_6^1 + S(4,2)C_6^2 2! + S(4,3)C_6^3 3! + S(4,4)C_6^4 4!$$
$$+ S(4,1)C_5^1 + S(4,2)C_5^2 2! + S(4,3)C_5^3 3! + S(4,4)C_5^4 4!$$
$$+ S(4,1)C_4^1 + S(4,2)C_4^2 2! + S(4,3)C_4^3 3! + S(4,4)C_4^4 4!$$
$$+ S(4,1)C_3^1 + S(4,2)C_5^2 2! + S(4,3)C_3^3 3!$$
$$+ S(4,1)C_2^1 + S(4,2)C_2^2 2!$$
$$+ S(4,1)C_1^1$$

c'est-à-dire :

$$7^4 + 6^4 + 5^4 + 4^4 + 3^4 + 2^4 + 1^4$$

8 - COROLLAIRE

$$\mathcal{V}(\bar{x}') = \frac{\sigma^2}{N-1}\left[\left(1-\frac{1}{N}\right)^{n-1} + \left(1-\frac{2}{N}\right)^{n-1} + \ldots + \left(1-\frac{N-1}{N}\right)^{n-1}\right]$$

$$= \frac{\sigma^2}{N-1}\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)^{n-1}$$

En effet :

$$\mathcal{V}(x') = \mathcal{E}\,[\mathcal{V}(x')|n_d] = \frac{N\sigma^2}{N-1}\,(n_d^{-1}) - \frac{\sigma^2}{N-1}$$

$$= \frac{N\sigma^2}{N-1}\left[\frac{N^{n-1}}{N^{n-1}} + \left(1-\frac{1}{N}\right)^{n-1} + \left(1-\frac{2}{N}\right)^{n-1} + \ldots + \frac{1}{N^{n-1}}\right]\frac{1}{N} - \frac{\sigma^2}{N-1}$$

$$= \frac{\sigma^2}{N-1}\left[\left(1-\frac{1}{N}\right)^{n-1} + \left(1-\frac{2}{N}\right)^{n-1} + \ldots + \frac{1}{N^{n-1}}\right]$$

Autrement dit :

$$\mathcal{V}(\bar{x}') = \frac{\sigma^2}{N-1}\,\frac{(N-1)^{n-1} + (N-2)^{n-1} + \ldots + 2^{n-1} + 1}{N^{n-1}}$$

9 - RESULTAT ASYMPTOTIQUE ($N \longrightarrow \infty$)

Théorème :

Si $N \longrightarrow \infty$, $\qquad \mathcal{V}(\bar{x}') \# \frac{\sigma^2}{n}$

En effet $\mathcal{V}(\bar{x}')/\sigma^2$ est la moyenne :

$$\frac{1}{N-1}\left[\left(1-\frac{1}{N}\right)^{n-1} + \left(1-\frac{2}{N}\right)^{n-1} + \ldots + \left(1-\frac{N-1}{N}\right)^{n-1}\right]$$

En première approximation, c'est

$$\int_0^1 (1-u)^{n-1}\,du = \left[\frac{1}{n}\,(1-u)^n\right]_1^0 = \frac{1}{n}$$

Remarque : Si N est beaucoup plus grand que n ; on conçoit que n_d ne soit guère inférieur à n (sinon avec une probabilité très faible) : il faut un grand hasard pour tirer 2 fois la même unité en n coups. Le résultat s'explique donc.

Toutefois si n n'est pas très petit, la probabilité d'avoir $n_d = n$, soit

$$N(N-1)\,(N-2)\ldots(N-n+1)/N^n$$

est loin d'approcher 1, elle peut facilement tomber jusqu'à 0,50. Et il convient de ne pas confondre $\mathcal{V}(\bar{x}')$ et σ^2/n.

10 - EMPLOI DE LA METHODE DES TRAPEZES

L'aire $\int_0^1 (1-x)^{n-1}dx = 1/n$ peut être bornée par l'aire de trapèzes inscrits et circonscrits de hauteur $1/N$:

a) Trapèzes circonscrits :

$$\frac{1}{N}\left[\frac{0}{2} + \left(\frac{1}{N}\right)^{n-1} + \ldots + \left(\frac{N-1}{N}\right)^{n-1} + \frac{1}{2}\right] > \frac{1}{n}$$

b) Trapèzes inscrits :

$$\frac{1}{N}\left[\left(\frac{1}{N}\right)^{n-1}+\left(\frac{2}{N}\right)^{n-1}+\ldots+\left(\frac{N-1}{N}\right)^{n-1}\right]+\int_0^{1/2N}u^{n-1}du+\int_{1-1/2N}^{1}u^{n-1}\,du<\frac{1}{n}$$

Posons :

$$\mathcal{V}(\bar{x}') = \frac{N\sigma^2}{N-1}\,T,$$

on a donc :

$$\frac{1}{n}-\frac{1}{2N}<T<\frac{1}{n}-\int_0^{1/2N}u^{n-1}du-\int_{1-1/2N}^{1}u^{n-1}\,du=\int_{1/2N}^{1-1/2N}u^{n-1}\,du$$

$$=\frac{1}{n}\left[\left(1-\frac{1}{2N}\right)^n-\left(\frac{1}{2N}\right)^n\right],$$

Développons

$$\frac{1}{n}-\frac{1}{2N}<T<\frac{1}{n}-\frac{1}{2N}+\frac{n-1}{8N^2}-\frac{(n-1)(n-2)}{48N^3}+\ldots<\frac{1}{n}-\frac{1}{2N}+\frac{n-1}{8N^2}$$

11 - COMPARAISON AVEC LES SONDAGES EXHAUSTIF ET BERNOULLIEN

a) Le facteur de $N\sigma^2/(N-1)$ serait $(1/n - 1/N)$ pour le sondage exhaustif (au lieu de T).

L'accroissement est $\frac{1}{2N}\left(1-\frac{n-1}{4N}+\ldots\right)$, soit $\left(\frac{1}{2N}\right)$ environ

b) Pour le sondage bernoullien, il faut comparer T à $(N-1)/nN$:

$\frac{1}{n}-\frac{1}{nN}>\frac{1}{n}-\frac{1}{2N}$ est évident :

$$\frac{1}{n}-\frac{1}{nN}>\frac{1}{n}-\frac{1}{2N}+\frac{n-1}{2N}+\frac{n-1}{8N^2}\Longleftrightarrow\left(\frac{1}{2}-\frac{1}{n}\right)>\frac{n-1}{8N}$$

ce qui exige $n > 2$ (avec $N > n$)

Mais le cas $n = 2$ se traite à part sans difficulté. On a donc bien

$$\frac{1}{n}\left(\frac{N-1}{N}\right)>\frac{1}{n}-\frac{1}{2N}+\frac{n-1}{8N^2}>\frac{1}{n}\left[\left(1-\frac{1}{2N}\right)^n-\left(\frac{1}{2N}\right)^n\right]>T$$

donc

$$\mathcal{V}(\bar{x}) > \mathcal{V}(\bar{x}')$$
(bernoullien)

Le calcul ci-dessus a l'avantage de donner une limite au gain de variance réalisé :

$$\mathcal{V}\bar{x}-\mathcal{V}\bar{x}'>\frac{N\sigma^2}{N-1}\left(\frac{1}{2N}-\frac{1}{nN}-\frac{n-1}{8N^2}\right)=\frac{\sigma^2}{2(N-1)}\left[\frac{n-2}{n}-\frac{n-1}{4N}\right]$$

Si n est grand et N très grand , le crochet est de l'ordre de 1. Le sondage exhaustif ramène la variance de $\frac{\sigma^2}{n}$ vers $\sigma_1^2\left(\frac{1}{n} - \frac{1}{N}\right)$; l'estimateur $\overline{x}'$ la ramène vers $\sigma^2\left(\frac{1}{n} - \frac{1}{2N}\right)$; le gain en variance est d'environ 50 % (par rapport au sondage exhaustif).

12 - NOTE BIBLIOGRAPHIQUE ET HISTORIQUE :

1/ La distribution de probabilité de n_d est apparue assez tard dans la littérature des sondages. Nous étudions déjà le sondage bernoullien avec identification (c'est-à-dire $\overline{x}_d$) dans un document intérieur à l'I. S. E. E. E. - S. E. E. F. (déc. 1955) ; et nous le reprenons dans l'Etude Théorique N° 7 de l'I. N. S. E. E. (1957) pages 119-120 et dans notre thèse (1958) pages 408-409 [1.2]. Mais nous ignorions l'existence des nombres de Stirling(1). Nous écrivions (avec ν à la place de N) en 1957 :

	Boules distinctes		Probabilités
n = 3	$n_d = 3$		$\nu(\nu-1)(\nu-2)/\nu^3$
	2		$3\nu(\nu-1)/\nu^3$
	1		ν/ν^3
n = 4	$n_d = 4$		$\nu(\nu-1)(\nu-2)(\nu-3)/\nu^4$
	3		$6\nu(\nu-1)(\nu-2)/\nu^4$
	2	type aaab	$4\nu(\nu-1)/\nu^4$
		type aabb	$3\nu(\nu-1)/\nu^4$
	1		ν/ν^4

et nous ajoutions : Le dénombrement direct semble difficile à généraliser. En 1958 nous répétions les mêmes formules ainsi que les suivantes (trouvées en 1955).

$$a_2 = 2^{n-1} - 1$$
$$a_3 = (3^{n-1} - 2^n + 1)/2!$$
$$a_4 = (4^{n-1} - 3^n + 3.2^{n-1})/3! \quad \text{etc..}$$

qui ne sont autres que S(n, 2), S(n, 3), S(n, 4).

Nous donnions (1958) un argument fort et rapide pour justifier : $\mathcal{V}(\overline{x}) > \mathcal{V}(\overline{x}')$; à savoir que $\overline{x}'$ est la projection de $\overline{x}$ (dans un espace défini dans la thèse), d'où la formule de Pythagore :

$$\mathcal{V}\overline{x} = \mathcal{V}\overline{x}' + \mathcal{E}(\overline{x} - \overline{x}')^2$$

2/ Le problème consistant à établir que $\mathcal{V}(\overline{x}) > \mathcal{V}(\overline{x}')$ a fait l'objet de recherches considérables et très théoriques de la part de l'école indienne - RAO (1966) [3] mentionne BASU (1958) [4], DES RAJ et KHAMIS (1958) [5], ROY et CHAKRAVARTI (1960) [6]. Ce second article de 1958 fait usage des nombres de Stirling. Toutefois DES RAJ et KHAMIS s'en

---

(1) C'est RIORDAN qui par son livre de Combinatoire nous les a fait découvrir.

servent pour établir l'inégalité

$$\mathcal{V}(\bar{x}) > \mathcal{V}'(\bar{x}')$$

On ne trouve pas trace chez eux de l'expression exacte de $\mathcal{V}\bar{x}'$ ou de $\mathcal{E}(n_d^{-1})$. Cette formule est pourtant connue de RAO (1966) dont c'est la formule (3) et qui en donne même l'expression approchée (4) :

$$\left[1 - \frac{n}{2N} + \frac{n(n-1)}{12N^2}\right] \frac{S^2}{n} \quad , \quad \text{avec} \quad S^2 = \frac{N\sigma^2}{N-1}$$

Nous avions trouvé $8N^2$ comme limite supérieure (par les trapèzes inscrits) ; elle n'est pas incompatible avec le $12N^2$ de RAO comme moyenne. Malheureusement RAO ne donne aucune démonstration, ni aucune indication sur l'auteur des calculs. Les autres auteurs indiens publient des articles généralement extrêmement théoriques, étrangers à une question aussi terre à terre.

Cependant PATHAK (1964) [7] désigne par y_m la moyenne des m unités distinctes observées dans l'échantillon (page 800, ligne 2), donc appelle $\bar{y}_m$ notre $\bar{x}'$; mais il ne donne de $E(m^{-1})$ et de la variance que des expressions impraticables ou asymptotiques ; il ne fait aucun usage des nombres de Stirling.

Aux Etats-Unis, KOOP les emploie dans une communication au Congrès de l'Institute of Mathematical Statistics, Cambridge, Massach. (Mai 1963) ; mais son papier ne concerne pas exactement le même problème [8].

REFERENCES BIBLIOGRAPHIQUES :

[1] THIONET (P.) - Les pertes d'information en théorie des sondages (1957). Imp. Nat. - Etudes Théoriques INSEE. N° 7.

[2] THIONET (P.) - La perte d'information par sondage (Thèse 1958)- Publications de l'I.S.U.P. 1959.

[3] RAO (JNK.) - On the comparison of sampling with and without replacement - Revue de l'I.I.S. - 34 - 2 - 1966 - p. 125-128.

[4] BASU (D.) - On sampling with and without raplacement - SANKHYA - 20 - 1958 - p. 287-294.

[5] DES RAJ, KHAMIS (S.H.) - Some remarks on sampling with replacement - A.M.S. 29 - 1958 - p. 550-557.

[6] ROY (J.), CHAKRAVARTI (I.M.) - Estimating the mean of a finite population - A.M.S. 31 - 1960 - p. 392-398.

[7] PATHAK (P.K.) - Sufficiency in sampling theory - A.M.S. 35 - 2 - June 1964 - p. 795-808.

[8] KOOP (J.C.) - An unbiased ratio estimator in sampling with replacement with unequal and varying selection probabilities - (94^{th}. Meeting of the IMS, Cambridge, Massach. May 1963).

RESUME

Il s'agit d'une mise au point et non d'un travail d'exploration.

On commence par rappeler la définition des nombres de STIRLING et leurs règles de calcul.

On indique ensuite comment les nombres de 2e espèce s'introduisent dans l'expression d'une distribution de probabilités concernant les tirages de Bernoulli : - (celle du nombre X de boules distinctes tirées en n coups) - puis dans celle de l'estimateur sans biais $\overline{x}'$ de la moyenne de population $\overline{X}$.

$\overline{x}'$ est par définition la moyenne des unités-échantillons distinctes.

On étudie enfin la variance de $\overline{x}'$: son expression exacte ; sa valeur asymptotique (grands échantillons) ; une borne supérieure, sa position par rapport aux variances des sondages bernoulliens et exhaustifs.

Ces divers résultats avaient (en gros) déjà été trouvés indépendamment par divers spécialistes indiens et par l'auteur (1958) ; mais leur présentation systématique faisait défaut. De même les nombres de Stirling ont été peu employés jusqu'ici dans les problèmes de sondage.