Bulletin de l'APMEP n° 379 - juin 1991

Echanges

Simulation d'une loi normale: *applications aux sondages*

Daniel Saada

Rambouillet

Un sondage définit une variable aléatoire, dite hypergéométrique, dont la distribution est trop compliquée pour se prêter au calcul. En approchant cette loi par une loi binomiale, puis par une loi normale, on facilite ce calcul et la résolution des premiers problèmes que pose l'interprétation des sondages. Mais en formulant d'autres problèmes (vice sans fin!), on bute rapidement sur l'écriture intégrale de la loi normale. On peut alors (on doit même parfois) recourir à des simulations, lesquelles donneront une réponse numérique au problème, ce qui permettra d'attendre sereinement la solution théorique (si elle existe) et de la vérifier.

A- Les hypothèses et conventions utilisées

Parmi les N électeurs d'un pays donné, N_1 sont franchement satisfaits du Président, tous les autres étant mécontents ou indécis.

On interroge n électeurs au hasard : le nombre X de partisans du Président est une variable aléatoire qui prend les valeurs 0, 1, ..., n (on suppose $n \leq N_1$)

La probabilité de l'événement "$X = k$" sera notée pr($X = k$). Il est usuel que X soit arrondi et présenté en pourcentage ; ainsi pour $n = 1000$ et $X = 432$, il sera affirmé que $X = 43\%$.

L'événement "$X = 43\%$" s'explicite donc par "$425 \leq X \leq 434$", et pr($X = 43\%$) vaudra alors $\sum_{425}^{434} \text{pr}\,(X = k)$.

Le lecteur pourra, bien sûr, faire d'autres conventions d'arrondis. Il lui sera loisible aussi de faire varier à son gré les paramètres n, N, N_1 qui, sauf mention contraire, ont été fixés à $n = 1000$, $N = 10^7$, $N_1 = 4 \times 10^6$.

B- La loi hypergéométrique.

On sait que pr($X = k$) vaut $C_{N_1}^{k} \times C_{N-N_1}^{n-k} / C_N^n$, expression composée d'entiers vertigineux, que nous limiterons comme annoncé à

$$C_{4\times10^6}^{k} \times C_{6\times10^6}^{1000-k} / C_{10^7}^{1000}.$$

Indiquons trois méthodes de calcul possibles pour cette distribution, en nous bornant à sa valeur modale $k = 400$.

1. Calcul de pr(X = 400) par logarithmes.

C'est la méthode la plus naturelle qui vient à l'esprit pour contourner les dépassements de capacité, mais sa mise en œuvre est longue et engendre curieusement une perte non négligeable, mais non dirimante, de chiffres significatifs (comme on le verra plus tard). Nous avons donc :

$$\text{pr}\,(X = 400) = C_{4\times10^6}^{400} \times C_{6\times10^6}^{600} / C_{10^7}^{400}$$

$$= C_{1000}^{400} \times \prod_{i=0}^{399} \frac{4\times10^6 - i}{10^7 - i} \times \prod_{i=0}^{599} \frac{6\times10^6 - i}{10^7 - 400 - i}$$

avec $C_{1000}^{400} = \prod_{i=1}^{400} (1 + 600/i)$. La programmation par logarithme est donc possible et donne : $\boxed{\text{pr}(X = 400) = 2{,}574\ 610\ 568\ \%}$

2- *Calcul de pr (X = 400) par la formule de Stirling.*

Bien qu'elle procède par approximation, cette méthode se révèle aussi précise que la méthode logarithmique, tout en étant incomparablement plus rapide.

Nous partons de l'équivalent $n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\ e^{\frac{1}{12n}}$ qui fournit en plus une excellente valeur relative de $n!$ (voir [1], pages 25 à 30). On en déduit :

$$C_n^p = \frac{1}{\sqrt{2\pi n}} \left(\frac{n}{p}\right)^{p+\frac{1}{2}} \left(\frac{n}{n-p}\right)^{n-p+\frac{1}{2}} \exp\left(\frac{1}{12}\left(\frac{1}{n} - \frac{1}{p} - \frac{1}{n-p}\right)\right) \text{ et}$$

$$\operatorname{Ln}\left(C_n^p\right) = -\frac{1}{2}\operatorname{Ln}(2\pi n) + \left(p+\frac{1}{2}\right)\operatorname{Ln}\left(\frac{n}{p}\right) + \left(n-p+\frac{1}{2}\right)\operatorname{Ln}\left(\frac{n}{n-p}\right) + \frac{1}{12}\left(\frac{1}{n} - \frac{1}{p} - \frac{1}{n-p}\right)$$

Il suffit donc de programmer cette fonction de deux variables, puis de donner aux couples (n,p) les valeurs désirées. On trouve

$$\boxed{\text{pr}(X = 400) = 2{,}574\ 620\ 566\ \%}$$

ce qui est remarquable.

3- *Calcul de pr(X = 400) par multiplications.*

Cette méthode, la plus précise, nous servira de référence.

Si votre instrument de calcul peut enregistrer des réels positifs inférieurs à 10^{-100} ou supérieurs à 10^{100}, programmez telle quelle l'écriture de pr(X = 400) vue en 1. Vous obtiendrez :

$$\boxed{\text{pr}(X = 400) = 2{,}574\ 610\ 674\ \%}$$

Dans le cas contraire, voici un algorithme pour dégager les chiffres significatifs d'un produit $\prod_1^n d_i$ quand les d_i sont tous entre 0 et 1 :

```
0 -   P = 1
1 -   pour i allant de 1 à n
2 -   P = P * d_i
3 -   si P < 0,1  faire P = 10*P
      et retourner en 3.
```

On a ainsi contraint les produits partiels à rester dans l'intervalle [1/10 , 1]. Quand les d_i dépassent 1, on adapte la ligne 3 de l'algorithme :

```
3 bis- si P > 10 faire P = P/10
       et retourner en 3 bis.
```

Vous pourrez ainsi calculer, par exemple, les premier chiffres de $n!$ *pour tout* n avec une exactitude maximale.

4- Fonction de répartition de la loi X.

On utilise la valeur de pr($X = 400$) et la relation de récurrence :

$\dfrac{\text{pr}\,(X = k+1)}{\text{pr}\,(X = k)} = \dfrac{N_1 - k}{N - N_1 - n + k + 1} \times \dfrac{n - k}{k + 1}$ avec, répétons-le $N = 10^7$, $N_1 = 4 \times 10^6$ et $n = 1000$.

Deux boucles en sens contraire donnent :

$$\boxed{\text{pr}(X = 40\%) = \text{pr}(395 \leq X \leq 404) \approx 25{,}304\%}$$

Trois fois sur quatre, le sondage exprimé en pourcentage différera d'au moins un point avec la réalité (dans les conditions fixées par l'énoncé).

C- L'approximation binomiale.

Supposons que l'institut de sondage tire, au hasard toujours, mais *avec remise* cette fois, les numéros de Sécurité Sociale des personnes à interroger. La loi de X prend alors une forme beaucoup plus maniable :

$$\text{pr}\,(X = k) = \mathbf{C}_n^k \left(\frac{N_1}{N}\right)^k \left(1 - \frac{N_1}{N}\right)^{n-k}$$

Seule compte la proportion de partisans du Président ; comme la probabilité de réinterroger la même personne est très faible (dans les conditions usuelles), il est légitime d'espérer que cette loi binomiale va approcher d'assez près la loi hypergéométrique d'origine (voir [3], pages 262 à 263).

De fait, on trouve les mêmes données :

$\text{pr}\,(X = 400) = \mathbf{C}_{1000}^{400} (0{,}4)^{400} (0{,}6)^{600} = 2{,}574\,48\%$ contre 2,574 61% en réalité, ce qui est une approximation digne d'éloges.

On remarquera (voir B.1) que, tous comptes faits, on a remplacé $\prod_{i=0}^{399} \frac{4 \times 10^6 - i}{10^7 - i}$ par $(0,4)^{40}$, approximation par excès, et $\prod_{i=0}^{599} \frac{4 \times 10^6 - i}{10^7 - 400 - i}$ par $(0,6)^{60}$, approximation par défaut. Il est juste de dire que l'approximation est d'autant moins bonne que k s'écarte de la valeur centrale $n\frac{N_1}{N}$. Cette dégradation va-t-elle empêcher les fonctions de répartition de coïncider ? Il n'en est rien, car les probabilités $\text{pr}(X = k)$ décroissent vers zéro et, somme toute, seules les erreurs *relatives* grandissent.

Vérifions :

la nouvelle relation de récurrence $\frac{\text{pr}(X = k+1)}{\text{pr}(X = k)} = \frac{N_1}{N - N_1} \times \frac{n-k}{k+1}$

donne $\boxed{\text{pr}(X = 40\%) = 25{,}3017\%}$ au lieu de 25,3041%!

Comparons de même $\text{pr}(|X - 40\%| \leq 1\%) = \text{pr}(385 \leq X \leq 414)$:

$\boxed{66{,}6930\%}$ par la loi binomiale

$\boxed{66{,}6946\%}$ par la loi hypergéométrique.

Un sondage sur trois s'écarte de la réalité d'au moins 2 points : plaignons ceux qui glosent sur une telle variation!

D- Le modèle normal de Laplace-Gauss.

Quand le nombre n de sondés n'est pas trop petit, et que la proportion p de fidèles du Président n'est ni trop faible ni écrasante, on peut approcher la loi binomiale X par la loi normale Y de même moyenne (np) et de même variance $(np\,(1-p))$ que X (Voir [1] pages 135 à 143).

Ces conditions sont largement réalisées quand $n = 1000$ et $p = 0{,}4$. La loi X étant discrète, on estimera $\text{pr}(X = k)$ par $\text{pr}(k - \frac{1}{2} \leq Y \leq k + \frac{1}{2})$. Rappelons aussi que Z définie par $Z = \frac{Y - np}{\sqrt{np(1-p)}}$ suit une loi normale, de moyenne 0 et de variance 1 (on dit centrée réduite).

Donnons quelques exemples d'utilisation de la loi normale.

1- Calculs de pr(X = 400) et pr(X = 40%).

$$\text{pr}(X = 400) \approx \text{pr}(399{,}5 \le Y \le 400{,}5) = pr\left(|Z| \le \frac{0{,}5}{\sqrt{240}}\right)$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^{\frac{0{,}5}{\sqrt{240}}} e^{-t^2/2}\,dt$$

intégrale que l'on évalue numériquement. On lit

$$\boxed{\text{pr}(X = 400) = 2{,}574\ 71\%}$$

La loi Y étant continue, il convient d'interpréter l'événement "$X = 40\%$" par $395 \le Y < 405$, ce qui fait disparaître l'asymétrie causée par la règle usuelle d'arrondi sur la variable entière X. On a alors

$$\boxed{\text{pr}(X = 40\%) = \frac{2}{\sqrt{2\pi}} \int_0^{\frac{5}{\sqrt{240}}} e^{-t^2/2}\,dt = 25{,}311\ 4\%}$$

Récapitulons :

	Loi hypergéométrique	*Loi binomiale*	*Loi normale*
pr(X=400)	2,574 61	2,574 48	2,574 71
pr(X=40%)	25,304 1	25,301 7	25,313 4
pr(|X-40%| ≤1%)	66,694 6	66,693 0	66,707 8

2- Moyenne de \|X - 40%\|.

Grâce à l'approximation normale, on établit très facilement la distribution *utile* de la variable entière $|X - 40\%|$:

	0	1	2	3	4	5
Probabilités (en millièmes)	253	414	226	83	20	3

soit en moyenne 1,2% (un programme serait le bienvenu).

3- Médiane de \|X - np\|.

Cherchons, en toute généralité, l'intervalle médian $[np - \mu\ ,\ np + \mu]$ en dehors duquel figurera inéluctablement un sondage sur deux.

Ce problème s'écrit $\text{pr}(|Y - np| \leq \mu) = \frac{1}{2}$ ou $\text{pr}(|Z| \leq \frac{\mu}{\sqrt{npq}}) = \frac{1}{2}$.

D'où l'équation $\frac{1}{\sqrt{2\pi}} \int_0^{\frac{\mu}{\sqrt{npq}}} e^{-t^2/2}\, dt = \frac{1}{4}$, dont une solution est $\frac{\mu}{\sqrt{npq}} = 0{,}674\,490$. Le problème est donc résolu!

Avec $n = 1000$ et $p = 0{,}4$: $\mu \approx 10{,}5$, donc, plus d'un sondage sur deux sera en dehors de la fourchette [390,410].

4- Moyenne de $|X - np|$.

Utilisons la relation $|Y - np| = |Z| \sqrt{np(1-p)}$:

$$E(|Y - np|) = \sqrt{np(1-p)}\, E(|Z|).$$

Or, $\text{pr}(|Z|) \leq k) = \frac{2}{\sqrt{2\pi}} \int_0^k e^{-t^2/2}\, dt$, $k \geq 0$; la densité de $|Z|$ est donc la fonction $k \to \frac{2}{\sqrt{2\pi}} e^{-k^2/2}$ et sa moyenne vaut alors :

$$\frac{2}{\sqrt{2\pi}} \int_0^{+\infty} k e^{-k^2/2}\, dk = \sqrt{\frac{2}{\pi}}.$$ D'où $E|Y - np| = \sqrt{\frac{2}{\pi}}\, \sqrt{np(1-p)}$

C'est la moyenne cherchée puisqu'on a assimilé X à Y. Avec $n = 1000$ et $p = 0{,}4$: $E(|X - 400|) \approx 12{,}4$.

Remarque :
quand $p = 0{,}5$ et n est pair, on sait exprimer la moyenne de $|X - np|$: $\frac{n}{2} C_n^{n/2} \left(\frac{1}{2}\right)^n$ (voir [4] page 144). Pour une généralisation, consulter [3], page 193, exercice 23.

E- Simulations d'un sondage.

On va solliciter évidemment massivement le générateur de nombres aléatoires de la machine. Cette fonction, que nous noterons RAN, est sensée produire des décimaux U uniformément distribués entre 0 et 1 et, en principe, mutuellement indépendants. Rappelons que $E(U) = \frac{1}{2}$ et que $V(U) = \frac{1}{12}$.

1- Simulation d'une loi binomiale.

Commençons par simuler la variable X de Bernoulli de paramètre p(pr($X = 1$) = p et pr($X = 0$) = $1 - p$) :

```
1. R = RAN
2. si R ≤ p alors X = 1
3. Si R > p alors X = 0
```

Pour engendrer la variable binomiale X de paramètres n et p, on exécute n fois l'algorithme précédent :

```
0. Se donner n et p
1. X = 0
2. pour i allant de 1 à n
3. si RAN < p alors X = X + 1
```

Comme nous aurons besoin de plusieurs centaines de simulations, la ligne 3. devrait être lue plusieurs *centaines de milliers* de fois (quand $n = 1000$). D'où un temps de calcul prohibitif, que l'on va réduire drastiquement en simulant la loi normale approchant X.

2. Simulation approchée d'une loi normale centrée réduite.

Soit U_i une suite créée par RAN.

Nous savons que $V_n = \frac{1}{n}\sum_{1}^{n} U_i$, à valeurs dans [0,1], est de moyenne $\frac{1}{2}$ et de variance $\sqrt{\frac{1}{12n}}$; $W_n = \sqrt{12n}\,(V_n - 0{,}5)$ est donc centrée réduite.

Le théorème de la limite centrée (voir [3], pages 265 à 268) affirme que W_n converge vers la loi normale centrée réduite. Il était d'usage autrefois de choisir par commodité $n = 12$, car $W_{12} = \sum_{1}^{12} X_i - 6$, mais il est probable

que cette valeur de n est un peu faible. Sur une calculatrice, il y aurait intérêt à doubler n (au moins).

3- Simulation exacte d'une loi normale centrée réduite.

Un générateur remarquablement simple de lois normales a été découvert dans les années cinquante par BOX, MULLER et MARSAGLIA (voir [2] page 117) : $Y = \cos(2\pi U).\sqrt{-2\ln(X)}$ est normale centrée réduite quand U et V sont uniformes et indépendantes sur [0,1].
De plus $Z = \sin(2\pi U).\sqrt{-2\ln(X)}$ est normale et indépendante de Y.

L'algorithme suivant calcule Y et Z *sans* recourir aux fonctions circulaires et, de ce fait, est un peu plus rapide (en moyenne!) que la transposition des formules :

1. $U = 2*RAN - 1$ et $V = 2*RAN - 1$
2. si $U^2+V^2>1$ alors retourner en 1
3. $S=U^2+V^2$ $\quad M = \sqrt{\dfrac{-2\ln(S)}{S}}$
4. $Y = U * M$ et $Z = V * M$

Pour établir les formules de BOX, MULLET et MARSAGLIA, raisonnons sur deux lois normales Y et Z indépendantes, centrées réduites et posons $Y = \rho\cos\theta$ et $Z = \rho\sin\theta$. Il faut alors prouver :
a) que $U = \theta/2\pi$ est uniforme sur [0,1],
b) que $V = e^{-\rho^2/2}$ est uniforme sur [0,1]
c) que U et V sont indépendantes.

A titre d'exemple, démontrons le point b) :

$$\text{pr}\left(\rho^2 \le k^2\right) = \text{pr}\left(Y^2 + Z^2 \le k^2\right) = \frac{1}{2\pi}\int_{-k}^{k} e^{-y^2/2}\,dy \int_{-\sqrt{k^2-y^2}}^{\sqrt{k^2-y^2}} e^{-z^2/2}\,dz$$

car Y et Z sont indépendantes.
D'où $\text{pr}\left(\rho^2 \le k^2\right) = \dfrac{1}{2\pi}\iint_D e^{-(y^2+z^2)/2}\,dy\,dz$, D étant le disque de rayon k centré sur l'origine.

En intégrant à rayon constant r : $\text{pr}\left(\rho^2 \leq k^2\right) = \int_0^k r e^{-r^2/2} dr = 1 - e^{-k^2/2}$

ce qu'il fallait démontrer.

4- Formulaire :

Simulation d'une loi normale Y de moyenne m et d'écart-type σ :

$$Y = m + \sigma \cos(2\pi U) \cdot \sqrt{-2 \ln (V)} .$$

Simulation d'une loi binomiale X de paramètres n et p :

$$X = \text{int}\left(m + \frac{1}{2} + \sigma \cos(2\pi U) \cdot \sqrt{-2 \ln (X)}\right)$$

$$\text{avec } m = np \text{ et } \sigma = \sqrt{np(1 - p)}.$$

Simulation d'un sondage X arrondi au pourcentage entier :

$$X = \text{int}\left(\frac{m}{10} + \frac{1}{2} + \frac{\sigma}{10} \cos(2\pi U) \cdot \sqrt{-2 \ln (X)}\right).$$

F- Exemples de simulation

Commençons par tester notre nouvel outil sur un problème dont on connait la solution.

1- Moyenne de $|X - 400|$

On va fabriquer des lots de 100 sondages $X_1, X_2, \ldots X_{100}$ (avec n= 1000 et p = 0,4), faire la somme E des écarts $|X_i - 400|$ et afficher $E/100$.

Programme en BASIC :

```
05   MODE DEGRE
10   E = 0  S = √240   (pour accélérer
                        la ligne 24)
20   FOR  I = 1  TO  100
     22   X = cos(360*RAN)*√-2*LN(RAN)
     24   X = INT(400,5 + X*S)
     26   E = E + |X - 400|
     28   NEXT I
30   PRINT E/100
```

J'ai obtenu successivement : 12,45 ; 12,26 ; 11,39 ; 13,45 ; 12,79 , soit 12,49 en moyenne. Rappelons la valeur théorique : 12,36 (voir D.4)

2- Ecarts entre deux sondages :

Deux instituts organisent la même semaine deux sondages indépendants et de même taille : à quel écart en moyenne doit-on s'attendre entre leurs résultats X_1 et X_2 ?

Réponse sur 500 simulations ($n = 1000$, $p = 0{,}4$) :

$$\boxed{E(|X_1 - X_2|) \approx 17{,}52}$$

Une étude par loi normale donne $E(|X_1 - X_2|) = \frac{2}{\sqrt{\pi}}\sqrt{240} = 17{,}48$

(écrire $\sup(X_1, X_2) = \frac{1}{2}(X_1 + X_2 + |X_1 - X_2|)$ et calculer la moyenne du sup après avoir déterminé sa densité).

Deuxième moyen de contrôle : quand $p = \frac{1}{2}$, on sait exprimer $E(|X_1 - X_2|)$ sans l'approximation normale ([4] page 145).

Exercice : obtenir, par simulation, la distribution de $|X_1 - X_2|$ et estimer sa médiane.

3- Statistiques de rang.

Soit $X_1, X_2, \ldots X_k$ des sondages *indépendants* effectués durant la même semaine. En rangeant les k résultats par ordre croissant, on obtient de nouvelles variables aléatoires que nous noterons encore $X_1, X_2, \ldots, X_k$, avec donc : $X_1 \leq X_2 \leq \ldots \leq X_k$. Si leurs densités sont assez faciles à dégager, le calcul des moyennes est délicat.

Choisissons $k = 3$ et opérons par simulation pour estimer les espérances $E_i = E(X_j)$:

Etape 0 : préparation de l'expérience.

```
0.1  réserver X(J) pour J allant de 1 à k.
0.2  réserver et annuler E(J).
0.3  pour i allant de 1 à 100.
```

Etape 1 : fabrication d'un lot de k sondages indépendants

```
1.1  pour j allant de 1 à k
     X(J) = int(400,5+√240*cos(360*RAN)
                              *√(- 2 * LN(RAN)))
```

Etape 2 : classement des X(*J*) pour $k = 3$, on peut se contenter de :

```
2.1  si X(3) < X(1)   échanger X(3) et X(1)
2.2  si X(2) > X(3)   échanger X(3) et X(2)
2.3  si X(2) < X(1)   échanger X(2) et X(1).
```

Etape 3 : calcul des moyennes.

```
3.1  pour j allant de 1 à k :E(j)=E(j)+X(j).
```

Etape 4 : affichage des résultats.

Quand la boule sur i(0.3) est achevé :

```
4.1  pour j allant de 1 à k :
                          afficher E(J)/100.
```

Cinq cents simulations m'ont donné :

$$E(X_1) = 38{,}71\% \quad ; \quad E(X_2) = 40{,}07\% \quad ; \quad E(X_3) = 41{,}34\%.$$

On en déduit $E(X_3 - X_1) \approx 2{,}63\%$, ce qui laisse à penser que

$$E(\sup X_1 - \inf X_i) = \frac{k}{\sqrt{\pi}}\sqrt{240}$$

généralisant ainsi le résultat de F-2. Le lecteur intéressé pourra se reporter à l'ouvrage suivant (*) : *ORDER STATISTICS* ; H.A.DAVID ; John Wiley 1970 ; chapitre 3, paragraphe 2 et exercice 3.1.3.

Exercice : estimer la moyenne du plus petit écart, à savoir $\inf(X_2 - X_1, X_3 - X_2)$.

4- Probabilités conditionnelles.

Quelle est la probabilité qu'une évolution de l'opinion ne soit pas détectée par les sondages ? C'est une question importante mais difficile (voir [4], page 145) que nous traiterons sur un exemple.

(*) Référence aimablement communiquée par Paul-Louis HENNEQUIN.

Supposons donc qu'un sondage, effectué quand la proportion d'électeurs contents du Président est 0,4, donne X_1 partisans sur 1 000 sondés. Un mois après, le Président a renforcé ses positions : la fraction de la population qui lui est acquise est passée de $p_1 = 0{,}40$ à $p_2 = 0{,}42$; un nouveau sondage recueille alors X_2 avis favorables sur 1 000. Calculons par simulation $\text{pr}(X_2 \leq X_2)$. Cinq fois 100 simulations m'ont donné : 11, 17, 26, 21, 23, soit 19,6% en moyenne.

Une fois sur cinq, la variation de deux points de la cote présidentielle ne sera pas décelée. Lorsque X_1 et X_2 sont arrondis en pourcentage entier, cette probabilité grimpe à 23,4%. Il serait évidemment passionnant de calculer

$\text{pr}(X_2 \leq X_1$, sachant que $p_2 \geq p_1)$

et $\text{pr}(p_2 \geq p_1$, sachant que $X_2 \geq X_1)$ mais ce serait entrer là dans une autre problématique.

BIBLIOGRAPHIE

[1] *Initiation aux Probabilités.* L.GUERBER et P.-L.HENNEQUIN. Brochure APMEP.

[2] *The art of computer programming.*D.E.KNUTH, vol.2, Addison-Wesley.

[3] *Cours de Probabilités et Statistiques.*C.LEBŒUF, J.L.ROQUE, J.GUEGAND. Ellipses.

[4] *Mathématiques et Programmation.* D.SAADA. Belin.

Et pour en savoir plus sur la mathématique des sondages :

[5] *Théorie des sondages.*C.GOURIEROUX. Economica.