

# Respecter l'anonymat

Jean-Claude Régnier  
Université Lumière de Lyon 2

## Résumé:

Les instruments de raisonnement humain fournis par la statistique et le calcul des probabilités peuvent parfois nous rendre de grand service dans des travaux d'investigation lorsque ceux-ci reposent sur des données recueillies au moyen de questions adressées directement à des personnes dont il importe absolument de préserver l'anonymat. Par quels dispositifs est-il possible d'accéder à des informations inscrites dans le jardin secret des individus ? Comment peut-on procéder pour respecter ce secret ?

L'article qui suit se propose de reprendre les démarches de Warner<sup>1</sup> (1965) traitant de cette question et de les mettre en œuvre dans le domaine d'investigations en Sciences de l'Éducation.

## Introduction

Nombre de problématiques abordées dans le cadre des Sciences de l'Éducation peuvent ou pourraient s'appuyer sur des données appartenant au jardin secret des individus concernés par les investigations. Nous pouvons aussi penser que la multiplication des sondages, les propos qu'ils induisent dans les médias, la corrélation plus ou moins implicite qui est établie avec l'informatique et sa puissance de stockage et de traitement d'informations privées mettent ces individus dans un état de méfiance préjudiciable quant à la fiabilité des données requises.

Le législateur a lui-même établi des garde-fous au travers d'une législation fondée sur les deux thèmes "informatique" et "liberté".

Ainsi est-il pertinent d'imaginer des dispositifs permettant de recueillir des informations en préservant de manière absolue l'anonymat des personnes interrogées et de leur associer des traitements statistiques adéquats.

Identifions quelques thèmes où cet anonymat est un préalable absolu:

- la fraude aux examens universitaires,
- l'usage de "drogue" lors de la préparation d'examens universitaires,
- la pratique du vol de documents, de livres dans les centres de documentation,
- le "piratage" informatique,
- la mise au point de logiciels "virus" et les stratégies élaborées pour leur propagation,
- l'effet des campagnes d'information relatives à la question du SIDA

---

<sup>1</sup> Lebart et alii (1982) *Traitement des données statistiques* Dunod p.74 -76

Cet ouvrage rapporte que S.L. Warner en 1965 propose une stratégie dans un article intitulé "Randomized Response: a Survey Technique for Eliminating Evasive Answer Bias" J.A.S.A. vol 60 pp 63-69

Un numéro de «International Statistic Review» (août 1976) contient les communications relatives à cette question lors d'un congrès à Varsovie en septembre 1975.

Tous ces thèmes peuvent donner lieu à des problématiques inscrites dans le cadre des Sciences de l'Education. Or force est de constater qu'elles impliquent la prise en compte de données intimes aux individus et requérant la préservation du secret.

### **Un exemple pour introduire la méthode.**

A titre d'exemple pour exposer la procédure choisie, nous optons pour une étude du comportement de fraude lors d'examens universitaires.

Un enseignant ayant eu l'écho de quelques rumeurs sur des pratiques de copiages lors d'examens universitaires pourra avoir envie d'évaluer la part de réalité de ce fait. Il pourra s'intéresser à l'occurrence de ces pratiques, aux stratégies employées, aux représentations qui lui sont associées et surtout à la motivation qui détermine le recours à la fraude ainsi qu'à la prise de conscience des divers risques encourus.

Pour la clarté de cet exposé, nous nous réduirons à la question:

*Avez-vous fraudé au moins une fois aux examens de l'année universitaire précédente?*

La question initiale appelle alors deux modalités de réponse:

A<sub>1</sub> = « *j'ai fraudé au moins une fois* »

A<sub>2</sub> = « *je n'ai jamais fraudé* »

Nous excluons ici l'hypothèse d'un comportement de mensonge systématique. Nous proposons le protocole suivant:

« Je souhaite recueillir une information relative à la fraude aux examens universitaires pour infirmer ou confirmer des rumeurs à ce propos.

*Avez-vous fraudé au moins une fois aux examens de l'année universitaire précédente?*

oui

non

Toutefois pour préserver le secret total de votre réponse, je vous propose la démarche suivante:

Voici une urne contenant 20 boules numérotées de 1 à 20.

Dans l'isoloir, vous tirez une boule, vous prenez connaissance de son numéro et vous la remettez dans l'urne.

Si le numéro tiré est inférieur ou égal à 5, je vous demande de répondre par OUI (c'est vrai) ou NON (c'est faux) à l'assertion

A<sub>1</sub> = « *j'ai fraudé au moins une fois* »

Si le numéro tiré est supérieur ou égal à 6, je vous demande de répondre par OUI (c'est vrai) ou NON (c'est faux) à l'assertion

A<sub>2</sub> = « *je n'ai jamais fraudé* »

Mentionnez votre réponse sur le petit carton mis à votre disposition et mettez ce carton dans la seconde urne.

La constitution de l'échantillon des étudiants à interroger soulève un problème concret de faisabilité: on peut procéder à un tirage aléatoire<sup>2</sup> de plus d'une centaine d'étudiants parmi ceux de l'université concernée à partir des numéros d'inscription et les inviter par courrier à venir répondre à l'enquête à la manière des consultations électorales.

Comment pourrions-nous interpréter les abstentions, c'est à dire ceux qui ne viendront pas s'exprimer ?

Comment pouvons nous modéliser ce paramètre lié au comportement qui se traduit par venir et s'exprimer ou ne pas venir ?

Quel biais peut introduire ce paramètre ?

Par cette démarche, nous recueillons un ensemble de  $n$  bulletins portant les réponses OUI ou NON sans savoir si elles se rapportent à  $A_1$  ou à  $A_2$ .

Le secret du chacun est donc respecté.

### **Recourir au raisonnement probabiliste.**

C'est alors que le recours à un raisonnement probabiliste va nous aider à obtenir une estimation de la proportion *des étudiants ayant fraudé au moins une fois*.

Cette fois nous devons introduire un peu plus de formalisme mathématique en espérant ne pas induire un découragement chez le lecteur non initié.

<i>notation</i>	<i>information</i>	<i>état</i>
$\pi$	proportion réelle des étudiants ayant fraudé au moins une fois dans la population	inconnue
$\theta$	probabilité d'être invité à répondre à $A_1$	connue <sup>3</sup>
$1 - \theta$	probabilité d'être invité à répondre à $A_2$	connue
$\rho$	probabilité que l'étudiant interrogé ait répondu OUI à l'une ou l'autre des assertions $A_1$ ou $A_2$	inconnue

**Tableau 1 : codage des paramètres**

<sup>2</sup> à l'aide d'une table de nombres au hasard

<sup>3</sup> La valeur est déterminée par la procédure aléatoire choisie: par exemple la composition de l'urne selon la proportion des tirages favorables à  $A_1$  et le mode de tirage selon qu'il s'agit d'un tirage avec ou sans remise.

<i>notation</i>	<i>événement</i>	<i>probabilité</i>
A <sub>1</sub>	répondre à l'assertion A <sub>1</sub>	P(A <sub>1</sub> )= θ
A <sub>2</sub>	répondre à l'assertion A <sub>2</sub>	P(A <sub>2</sub> )= 1- θ
«OUI»	répondre OUI à A <sub>1</sub> ou A <sub>2</sub>	P(«OUI»)= ρ
(«OUI»/A <sub>1</sub> )	répondre OUI sachant qu'il s'agit de répondre à A <sub>1</sub>	P(«OUI»/A <sub>1</sub> )= π
(«OUI»/A <sub>2</sub> )	répondre OUI sachant qu'il s'agit de répondre à A <sub>2</sub>	P(«OUI»/A <sub>2</sub> )= 1- π

**Tableau 2 : les évènements et leur mesure de probabilité**

L'évènement «OUI» est réalisé par les éventualités qui elles-même réalisent soit l'évènement «OUI» et A<sub>1</sub> soit l'évènement «OUI» et A<sub>2</sub>.

$$«OUI» = («OUI» et A_1) \text{ ou } («OUI» et A_2)$$

Ainsi en terme de probabilités:

$$P(«OUI») = P(«OUI» et A_1) + P(«OUI» et A_2)$$

Le recours à la notion de "probabilité conditionnelle" nous permet d'obtenir les relations suivantes:

$$P(«OUI» et A_1) = P(«OUI» / A_1) P(A_1)$$

$$P(«OUI» et A_2) = P(«OUI» / A_2) P(A_2)$$

$$P(«OUI») = P(«OUI» / A_1) P(A_1) + P(«OUI» / A_2) P(A_2)$$

$$P(«OUI») = \boxed{\pi\theta + (1-\pi)(1-\theta) = \rho} \quad \text{(Formule 1)}$$

En transformant cette relation nous extrayons une formule fournissant la valeur qui nous préoccupe à savoir celle de π

$$\boxed{\pi = \frac{\rho + \theta - 1}{2\theta - 1}} \quad \text{(Formule 2)}$$

Si maintenant nous tenons la variable "répondre à l'assertion" comme une variable aléatoire, nous avons alors à faire à une variable de Bernoulli puisqu'il n'y a que deux modalités «OUI» et «NON».

événements	«OUI»	«NON»
probabilité	ρ	1- ρ

**Tableau 3 : Deux évènements élémentaires et leur mesure de probabilité**

En interrogeant les n étudiants de l'échantillon, nous pouvons définir la variable aléatoire X qui décrit le nombre de "oui" émis parmi les n réponses.

X est une variable binomiale de paramètres n et ρ :  $X = B(n, \rho)$

Nous connaissons:

- l'espérance mathématique de X : E(X)= np

- la variance de X :  $\sigma^2(X) = V(X) = n\rho(1-\rho)$

Nous pouvons définir la variable aléatoire Y caractérisant la proportion de réponses "oui" dans l'échantillon.

$$Y = \frac{X}{n} \text{ avec } E(Y) = \rho \text{ et } \sigma^2(Y) = V(Y) = \frac{\rho(1-\rho)}{n} \quad \text{(Formule 3)}$$

Ceci nous conduit à procéder à une estimation ponctuelle de la valeur  $\rho$  par utilisation d'une variable aléatoire : la statistique  $\hat{\rho} = \frac{X}{n} = Y$  dont les caractéristiques sont  $E(\hat{\rho}) = \rho$  et  $V(\hat{\rho}) = \frac{\rho(1-\rho)}{n}$

A partir de cette relation, nous pouvons construire un estimateur  $\hat{\pi}$  de la valeur  $\pi$  inconnue.

$$\hat{\pi} = \frac{\hat{\rho} + \theta - 1}{2\theta - 1} \quad \text{(Formule 4)}$$

Nous connaissons là encore:

- l'espérance mathématique de  $\hat{\pi}$  :  $E(\hat{\pi}) = \frac{E(\hat{\rho}) + \theta - 1}{2\theta - 1} = \pi$

- la variance de  $\hat{\pi}$  :  $V(\hat{\pi}) = \frac{1}{(2\theta - 1)^2} V(\hat{\rho}) = \frac{1}{(2\theta - 1)^2} \left(\frac{\rho(1-\rho)}{n}\right)$

De cette relation, nous parvenons à une expression en fonction de  $\pi$ ,

$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{\theta(1-\theta)}{n(2\theta - 1)^2} \quad \text{(Formule 5)}$$

### Recourir à l'estimation statistique .

Dans notre exemple, nous supposons que notre échantillon est de  $n = 150$  individus extraits aléatoirement d'une population dont l'effectif total N est inconnu.

Le dépouillement a fourni les résultats suivants:

84 "oui" ont été dénombrés dans l'urne

<i>notation</i>	<i>événement</i>	<i>probabilité</i>
A <sub>1</sub>	répondre à l'assertion A <sub>1</sub>	$P(A_1) = \frac{5}{20}$
A <sub>2</sub>	répondre à l'assertion A <sub>2</sub>	$P(A_2) = \frac{15}{20}$
«OUI»	répondre OUI à A <sub>1</sub> ou A <sub>2</sub>	estimation ponctuelle $P(\text{«OUI»}) = \frac{84}{150}$

Tableau 4

Ceci nous donne donc une estimation ponctuelle de la proportion des étudiants *ayant fraudé au moins une fois* au sein de la population entière.

$$\text{estimation de } \pi = \frac{\frac{84}{150} + \frac{5}{20} - 1}{2 \frac{5}{20} - 1} = \frac{0,19}{0,5} = 0,38$$

**Ainsi il y aurait 38% d'étudiants ayant fraudé au moins une fois**

En raison de l'effectif important de l'échantillon, nous pouvons approcher la variable  $Z = \frac{\hat{\pi} - E(\hat{\pi})}{\sqrt{V(\hat{\pi})}} = \frac{\hat{\pi} - \pi}{\sqrt{V(\hat{\pi})}}$  par la variable de Laplace-Gauss LG(0;1).

La variance  $V(\hat{\pi})$  est inconnue mais nous pouvons l'estimer par sa valeur maximale correspondant à la valeur  $\pi = 0,5$ .

$$V(\hat{\pi}) \approx \frac{1}{4n} + \frac{\theta(1-\theta)}{n(2\theta-1)^2} = \frac{1}{600} + \frac{0,1875}{37,5} \approx 0,00666$$

$$\sigma(\hat{\pi}) \approx 0,0816$$

Enfin nous pouvons obtenir l'intervalle de confiance<sup>4</sup> à 95% suivant:

$$(0,38) - 1,96\sigma(\hat{\pi}) < \pi < (0,38) + 1,96\sigma(\hat{\pi})$$

c'est à dire:  $0,2199 < \pi < 0,5400$

**Ainsi peut on s'attendre avec une confiance de 95% à une proportion  $\pi$  d'étudiants ayant fraudé au moins une fois comprise entre 22% et 54%.**

Notons que la variance  $V(\hat{\pi})$  inconnue peut aussi être estimée par la valeur correspondant à la valeur  $\pi = 0,38$  issue de l'estimation ponctuelle.

$$V(\hat{\pi}) \approx \frac{1}{4n} + \frac{\theta(1-\theta)}{n(2\theta-1)^2} = \frac{0,2356}{150} + \frac{0,1875}{37,5} \approx 0,00657$$

$$\sigma(\hat{\pi}) \approx 0,0810$$

Ce qui nous permet d'obtenir l'intervalle de confiance à 95% suivant:

$$(0,38) - 1,96\sigma(\hat{\pi}) < \pi < (0,38) + 1,96\sigma(\hat{\pi})$$

c'est à dire:  $0,2211 < \pi < 0,5388$

### Question du choix de la valeur de $\theta$

Un autre problème se pose quant au choix de la valeur du paramètre  $\theta$ . Comment l'individu interrogé réagit-il selon la valeur du paramètre  $\theta$ ? Existe-t-il un seuil en deçà ou au-delà duquel le choix aléatoire de la modalité sur laquelle il doit se prononcer lui semble "truqué" c'est à dire qu'il lui paraît possible de voir son secret trahi ?

<sup>4</sup> une "fourchette" d'estimation de la proportion  $\pi$

Pour éclairer le sens de cette question, il convient d'analyser la situation sous deux contraintes antagonistes:

- d'une part plus la probabilité de tirer A<sub>1</sub> est proche de 0 (ou de 1), plus l'enquêteur recueille de réponses concentrées sur A<sub>2</sub> (ou sur A<sub>1</sub>)
- d'autre part plus la probabilité est proche de 1/2, plus le secret de l'enquête est préservé puisque A<sub>1</sub> et A<sub>2</sub> auraient les mêmes chances d'être tirées.

La situation  $\theta = 0,5$  ne peut être mise en œuvre puisque dans ce cas

$$P(\text{«OUI»}) = \pi \frac{1}{2} + (1 - \pi) \left(1 - \frac{1}{2}\right) = \rho = \frac{1}{2}$$

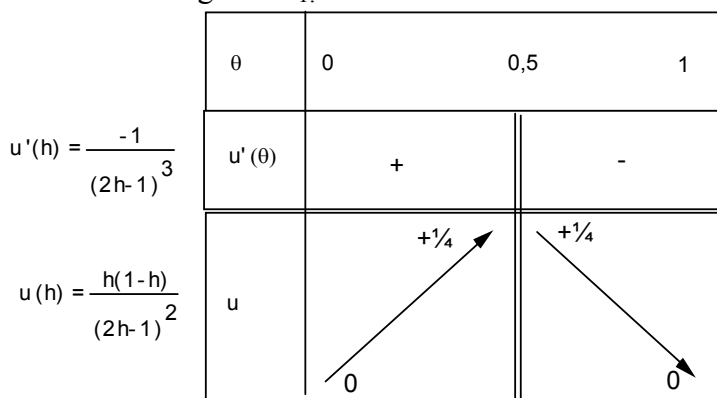
Cette probabilité serait indépendante de la proportion  $\pi$  que nous cherchons à estimer et la variance de l'estimateur  $\pi$  serait alors infinie. De ce fait aucune estimation ne pourrait être réalisée.

La situation  $\theta = 0$  ne peut être mise en œuvre puisque dans ce cas cela revient ne soumettre que la formulation A<sub>2</sub>

La situation  $\theta = 1$  ne peut être mise en œuvre puisque dans ce cas cela revient ne soumettre que la formulation A<sub>1</sub>

Ainsi il nous faut choisir une valeur de  $\theta$  sur l'ensemble  $]0 ; 0,5[ \cup ]0,5 ; 1 [$  Mais alors quels critères pouvons-nous retenir?

Nous pourrions choisir une valeur qui minimise autant que possible la variance de l'estimateur de  $\pi$ . tout en laissant l'impression que la dissymétrie dans le tirage entre A<sub>1</sub> et A<sub>2</sub> n'est pas trop déséquilibrée. Analysons la variation de la variance de l'estimateur de  $\pi$  en fonction de cette probabilité  $\theta$  de tirage de A<sub>1</sub>.



$$u'(\theta) = \frac{-1}{(2\theta - 1)^3}$$

$$u(\theta) = \frac{\theta(1 - \theta)}{(2\theta - 1)^2}$$

d'autre part nous pouvons noter  $w_n(\theta)$  la variance de l'estimateur de  $\pi$  en fonction de  $n$ ,  $\theta$  et  $h$

$$w_n(\theta) = \frac{\theta(1 - \theta)}{n} + \frac{1}{n} u(\theta)$$

$$w'_n(\theta) = \frac{1}{n} u'(\theta)$$

Nous fournissons en annexe la courbe représentative de la fonction  $u$  ainsi qu'un tableau des valeurs de  $u(\theta)$  selon diverses valeurs de  $\theta$

Quel choix peut être raisonnablement fait pour  $\theta$  du point de vue de l'étudiant interrogé ?

halshs-00405764, version 1 - 26 Jul 2009

Le seul moyen de parvenir à une estimation de cette valeur nous semble être de procéder à un sondage préalable mettant en évidence les réactions des individus interrogés face à la probabilité de répondre à  $A_1$  plutôt qu'à  $A_2$ .

Le problème peut se réduire en principe à estimer ce paramètre dans l'intervalle  $]0 ; 0,5 [$  si nous admettons qu'il revient au même de répondre non à  $A_1$  et oui à  $A_2$ , oui à  $A_1$  et non à  $A_2$ . Ce qui n'est peut-être qu'une hypothèse pratique.

Nous proposons l'expérience préalable suivante:

<p>« Je souhaite recueillir une information relative à la fraude aux examens universitaires pour infirmer ou confirmer des rumeurs à ce propos.</p> <p><i>Avez-vous fraudé au moins une fois aux examens de l'année universitaire précédente?</i></p> <p style="text-align: center;"><input type="checkbox"/> oui    <input type="checkbox"/> non</p> <p>Toutefois pour préserver le secret total de votre réponse, je vous propose la démarche suivante: Voici dix urnes contenant 100 boules portant chacune soit la marque <b>(A1)</b> soit la marque <b>(A2)</b>. Chaque urne est constituée d'une proportion différente de boules de deux types. Il s'agira de tirer une boule au hasard pour apporter une réponse à la modalité figurant sur cette boule. Regardez bien la composition de chaque urne avant d'en choisir une. Après quoi vous seriez amené à procéder comme suit: Dans l'isoloir, vous tirez une boule, vous prenez connaissance de son numéro et vous la remettez dans l'urne. Si la boule tirée porte la mention <math>A_1</math>, je vous demande de répondre par OUI (c'est vrai) ou NON (c'est faux) à l'assertion <math>A_1 = \text{« } j'ai fraudé au moins une fois \text{ »}</math> Si la boule tirée porte la mention <math>A_2</math>, je vous demande de répondre par OUI (c'est vrai) ou NON (c'est faux) à l'assertion <math>A_2 = \text{« } je n'ai jamais fraudé \text{ »}</math> Mentionnez votre réponse sur le petit carton mis à votre disposition et mettez ce carton dans la seconde urne.</p>
--

Il pourrait être intéressant de recueillir avec le premier choix de l'urne, un second sous la contrainte:

quelle urne contenant la plus petite quantité de boules portant la mention  $A_1$  accepteriez-vous de choisir?

Il est possible que la taille de l'urne joue un rôle dans l'impression laissée quant aux chances de réaliser un tirage le plus équitable possible entre  $A_1$  et  $A_2$ . Ainsi en proposant des urnes de 20 boules nous ne recueillerions peut-être pas les mêmes proportions choisies par le fait que rien n'assure qu'intuitivement pour un certain nombre de personnes non averties, il revient au même d'avoir 8 boules **(A1)** sur 20 que 40 boules **(A1)** sur 100!

Nous proposons de réaliser prochainement cette expérience dont nous communiquerons les résultats dans un prochain article. Dorénavant, il

appartient alors au lecteur d'exploiter cette information dans la conduite de quelques investigations soucieuses de respecter le secret absolu de l'individu interrogé. Nous nous en tiendrons là dans l'exploitation de ce dispositif de recueil de données. Nous espérons tout à la fois ne pas avoir rebuté le lecteur par le recours à une instrumentation mathématique et lui avoir fourni une matière à réflexion en ce qui concerne l'utilité de raisonnements statistiques et probabilistes. Nous visons aussi dans cet article une illustration de la statistique et des probabilités comme non pas des outils d'inquisition mais comme des outils soucieux du respect de l'intégrité, de l'intimité, du secret des sujets qui détiennent l'information indispensable aux délicates problématiques abordées.

**Annexe**

valeur de $\theta$	$u(\theta)$	valeur de $\theta$	$u(\theta)$	valeur de $\theta$	$u(\theta)$
0	0	0,17	0,32392103	0,34	2,19140625
0,005	0,00507601	0,175	0,34171598	0,345	2,35145682
0,01	0,0103082	0,18	0,36035156	0,35	2,52777778
0,015	0,01570305	0,185	0,37988158	0,355	2,72265161
0,02	0,02126736	0,19	0,4003642	0,36	2,93877551
0,025	0,02700831	0,195	0,4218624	0,365	3,17935528
0,03	0,03293345	0,2	0,44444444	0,37	3,44822485
0,035	0,03905076	0,205	0,46818443	0,375	3,75
0,04	0,04536862	0,21	0,4931629	0,38	4,09027778
0,045	0,05189591	0,215	0,51946753	0,385	4,47589792
0,05	0,05864198	0,22	0,54719388	0,39	4,91528926
0,055	0,06561672	0,225	0,57644628	0,395	5,41893424
0,06	0,07283058	0,23	0,60733882	0,4	6
0,065	0,08029462	0,235	0,63999644	0,405	6,67520776
0,07	0,08802055	0,24	0,67455621	0,41	7,46604938
0,075	0,09602076	0,245	0,71116878	0,415	8,40051903
0,08	0,10430839	0,25	0,75	0,42	9,515625
0,085	0,11289737	0,255	0,79123282	0,425	10,86111111
0,09	0,1218025	0,26	0,83506944	0,43	12,505102
0,095	0,13103948	0,265	0,88173382	0,435	14,5428994
0,1	0,140625	0,27	0,93147448	0,44	17,11111111
0,105	0,15057683	0,275	0,9845679	0,445	20,411157
0,11	0,16091387	0,28	1,04132231	0,45	24,75
0,115	0,17165627	0,285	1,10208221	0,455	30,6141975
0,12	0,18282548	0,29	1,16723356	0,46	38,8125
0,125	0,19444444	0,295	1,23720999	0,465	50,7704082
0,13	0,20653762	0,3	1,3125	0,47	69,1944444
0,135	0,21913117	0,305	1,39365549	0,475	99,75
0,14	0,23225309	0,31	1,48130194	0,48	156
0,145	0,24593335	0,315	1,57615047	0,485	277,527778
0,15	0,26020408	0,32	1,67901235	0,49	624,75
0,155	0,27509977	0,325	1,79081633	0,495	2499,75
0,16	0,29065744	0,33	1,91262976	0,49999	62500000
0,165	0,30691691	0,335	2,04568411	0,49999999	6,25E+14

