Association
mathématique
du Québec

Article

Combien y a-t-il de personnes vaccinées dans mon groupe ?

Christian Genest, James A. Hanley et Sahir Rai Bhatnagar
Université McGill, Montréal
christian.genest@mcgill.ca

Résumé

Les auteurs expliquent comment estimer le taux de couverture vaccinale d'un groupe sans interroger directement ses membres à ce propos. Chaque participant choisit au hasard une question à laquelle il peut répondre honnêtement et en toute liberté puisque personne d'autre que lui ne sait de laquelle il s'agit. Qui plus est, la participation du groupe est stimulée du fait que tous les répondants s'intéressent au résultat du sondage et qu'il est possible de le partager instantanément. L'exercice offre en prime la possibilité de s'initier à l'estimation et au calcul d'erreur statistique dans un cadre simple.

Mots clés : couverture vaccinale, incitation à la franchise, respect de la vie privée, sondage à question aléatoire

1 Introduction

Dans un message adressé en juillet 2021 à toute la communauté universitaire mcgilloise, un haut responsable des études et de la vie étudiante de notre établissement affirmait ce qui suit :

> « Au Québec, une personne en situation d'autorité n'a pas le droit de s'enquérir du statut vaccinal d'un subordonné : c'est illégal. Par conséquent, un supérieur ou un professeur ne peut pas demander à un employé ou à un étudiant s'il a été vacciné. »

Quoique cette consigne, maintes fois répétée par la suite, ne soit pas à l'abri d'éventuelles contestations devant les tribunaux, elle n'en a pas moins frustré de nombreux étudiant.e.s et membres du personnel inquiets pour leur sécurité au moment de se présenter sur le campus pour la première fois depuis le début de la pandémie, si ce n'est pour la première fois de leur vie.

Au moment d'écrire ces lignes, cette préoccupation se manifeste partout dans le réseau de l'éducation québécois, tant au secondaire qu'au collégial et à l'université. Or, bien que nous

disposions d'informations rassurantes quant au degré de couverture vaccinale au plan national, y compris par tranches d'âge, une question demeure d'actualité pour beaucoup de gens :

> « Combien y a-t-il de personnes (pleinement) vaccinées dans mon groupe ? »

Pour y répondre, on pourrait réaliser en classe un sondage anonyme par voie électronique. Toutefois, la question est assez délicate pour que les membres du groupe se méfient d'une telle procédure, même si seul un sommaire des réponses était révélé. Si elles n'ont pas l'absolue certitude que leur anonymat est bel et bien protégé, les personnes présentes risquent de ne pas répondre ou de mentir pour se prémunir contre tout sentiment réprobateur en se conformant à la norme sociale. Par ailleurs, il n'est pas sûr qu'un sondage de ce type soit lui-même conforme aux diktats administratifs puisque la question du statut vaccinal aurait quand même été posée.

L'objectif de cet article est de présenter les grandes lignes d'une autre stratégie d'enquête qui, tout en étant très simple, permet de répondre à la question dans le respect des exigences légales et du droit à la vie privée. Il s'agit d'une technique d'entrevue classique qui a fait ses preuves et dont l'origine remonte au regretté Stanley Warner [14], statisticien d'origine américaine qui a longtemps été professeur à l'Université York, à Toronto [13].

La méthode de la question aléatoire (MQA) consiste à inviter chaque individu sondé à répondre par oui ou non à une question qu'il aura lui-même choisie au hasard. Supposons plus particulièrement que l'une des deux questions concerne le statut vaccinal de la personne tandis que l'autre porte sur un sujet futile. Comme seul le participant sait à quelle question il répond, on ne peut pas déduire de sa réponse si celui-ci est vacciné ou pas. Cette personne peut donc se sentir tout à fait à l'aise de participer au sondage et d'y répondre honnêtement, d'autant plus qu'il lui importe de savoir combien de membres de son groupe sont vaccinés.

Une version élémentaire de la MQA est présentée au §2. Au §3, on a recours à des notions probabilistes relativement simples pour expliquer en quoi cette approche donne une estimation fiable de la couverture vaccinale et au §4, on s'intéresse au calcul de la marge d'erreur associée à cette estimation. On montre en outre au §5 comment il est possible de réduire cette marge d'erreur en répétant le sondage à quelques reprises.

La plupart des calculs présentés ici étant à la portée d'étudiants de niveau collégial, cette problématique offre une occasion d'observer, dans un cas concret qui suscite l'intérêt, l'effet du hasard et la façon dont il peut être contrôlé pour atteindre l'équilibre fragile mais indispensable qui doit être assuré entre la quête d'information et la protection de la vie privée.

Des compléments bibliographiques sur la MQA sont fournis au §6 et les annexes renferment certaines technicalités ainsi qu'un guide pratique pour la réalisation d'un sondage en ligne.

2 La méthode de la question aléatoire (MQA)

Introduite dès 1965 par Stanley Warner [14], la méthode d'enquête à question aléatoire se décline de bien des manières, comme le relatent Bellhouse [1] et Blair et coll. [2] dans leurs survols. Dans un souci de vulgarisation, nous nous limiterons ici à l'approche la plus simple et la plus intelligible possible, formulée dans le cadre précis de la question vaccinale.

Étant donné un groupe de taille N, on s'intéresse au nombre n de personnes qui sont (pleinement) vaccinées. Une version élémentaire de la MQA fait intervenir deux questions, celle qui nous intéresse et une autre qui ne porte pas à controverse.

Par exemple :

> **Question A :** Êtes-vous une personne (pleinement) vaccinée ? Oui ou non
> **Question B :** Avez-vous obtenu pile au jet d'une pièce de monnaie ? Oui ou non

Étant donné un nombre $p \in]0,1[$ convenu à l'avance et connu de tous, les participants au sondage sont alors conviés à procéder comme suit :

1. Choisir aléatoirement un nombre U dans l'intervalle $[0,1]$.
2. Si $U \leq p$, répondre à la Question A.
3. Si $U > p$, jeter une pièce de monnaie équilibrée et répondre à la Question B.

Le choix du nombre aléatoire U peut être réalisé entre autres au moyen d'`Excel` ou du gratuiciel statistique R. On peut en outre remplacer le jet d'une pièce de monnaie par la génération d'un second nombre aléatoire V dans l'intervalle $[0,1]$ et déclarer avoir obtenu « pile » si $V \leq 1/2$.

Cette procédure assure la protection du participant puisque lui seul sait à quelle question il a répondu. Avoir dit « oui » peut signifier qu'il est vacciné ou qu'il a obtenu pile. De même, avoir répondu « non » peut vouloir dire qu'il n'est pas vacciné ou qu'il a obtenu face.

Néanmoins, le nombre X de personnes qui répondent « oui » permet d'estimer le nombre inconnu n de personnes vaccinées dans le groupe. En effet, on démontrera au §3 qu'une estimation sans biais de n est donnée par la formule

$$\hat{n} = \frac{X - N(1-p)/2}{p}\,. \tag{2.1}$$

De plus, on verra aux §4 et §5 que la marge d'erreur déduite d'un intervalle de confiance asymptotique de niveau 95% est d'environ

$$\pm 2\sqrt{\frac{N(1-p)(1+p)}{4p^2}}$$

et que si l'on répète le sondage Q fois, cette marge d'erreur est réduite à

$$\pm\frac{2}{\sqrt{Q}}\sqrt{\frac{N(1-p)(1+p)}{4p^2}}\,.$$

Supposons par exemple que cette procédure soit appliquée Q fois sur un groupe de $N = 40$ personnes en prenant $p = 1/2$. La marge d'erreur est alors de

$$\pm\frac{2}{\sqrt{Q}}\sqrt{30} \approx \frac{1}{\sqrt{Q}}\,11.$$

Si on récolte $X = 24$ « oui » lors d'un seul sondage (c'est-à-dire $Q = 1$), l'estimation du nombre d'individus pleinement vaccinés est alors $\hat{n} = (24 - 10)/0{,}5 = 28$ et la marge d'erreur est de ± 11 personnes « 19 fois sur 20 ». En revanche, si la moyenne de « oui » est $\bar{X}_Q = 24$ sur $Q = 4$ répétitions, la marge d'erreur est réduite de moitié, soit $11/2 = 5{,}5$.

3 Pourquoi est-ce que ça fonctionne ?

L'estimateur $\hat{n}$, fondé sur la méthode des moments, est sans biais. En effet, soit $n \in \{0, \ldots, N\}$ le nombre inconnu mais fixe de membres du groupe qui sont pleinement vaccinés. Sans perte de généralité, la variable aléatoire X peut s'exprimer sous la forme

$$X = X_1 + \cdots + X_N$$

en termes de N variables de Bernoulli mutuellement indépendantes, où pour tout entier $i \in \{1, \ldots, N\}$, X_i vaut 1 si la réponse du i^{e} participant est « oui » et 0 autrement.

On peut supposer sans perte de généralité que les individus vaccinés correspondent aux indices $i \in \{1, \ldots, n\}$. Pour ces personnes, on a

$$X_i = 1 \quad \Leftrightarrow \quad U \leq p, \text{ ou encore } U > p \text{ et pile.}$$

Ainsi, pour tout entier $i \in \{1, \ldots, n\}$, on a

$$\Pr(X_i = 1) = p + (1-p)/2 = (1+p)/2.$$

De façon semblable, on vérifie que la probabilité qu'un individu $i \in \{n+1, \ldots, N\}$ non vacciné réponde « oui » est

$$\Pr(X_i = 1) = (1-p)/2.$$

Par linéarité de l'espérance, il s'ensuit que

$$\mathrm{E}(X) = n(1+p)/2 + (N-n)(1-p)/2 = np + N(1-p)/2.$$

Ainsi, l'estimateur de n fondé sur la méthode des moments est défini implicitement par l'équation

$$np + N(1-p)/2 = X,$$

où X est le nombre observé de « oui ». En isolant n dans cette équation, on trouve la formule de $\hat{n}$ donnée en (2.1). Cet estimateur est sans biais, puisque par linéarité de l'espérance, $\mathrm{E}(\hat{n}) = n$.

Le même raisonnement s'applique lorsque le sondage est répété Q fois de façon indépendante. Il suffit alors de remplacer X dans la formule (2.1) par $\bar{X}_Q = (X_1 + \cdots + X_Q)/Q$, c'est-à-dire le nombre moyen de « oui » obtenus sur l'ensemble des Q répétitions.

Comme on le verra au §5, l'un des bénéfices découlant de la répétition du sondage est qu'il est alors possible de réduire la variabilité de l'estimateur $\hat{n}$. De fait, il découle des lois faible et forte des grands nombres que $\bar{X}_Q$ converge en probabilité et presque sûrement vers $\mathrm{E}(X)$ quand $Q \to \infty$. Il s'ensuit que l'estimateur $\hat{n}$ est convergent, c'est-à-dire que $\hat{n} \to n$ quand $Q \to \infty$.

Voilà deux propriétés rassurantes qui s'ajoutent à la simplicité de l'estimateur $\hat{n}$ et qui militent en sa faveur. Celui-ci n'est pas parfait pour autant. On peut noter en particulier que $\hat{n}$ n'est pas forcément un entier et qu'au surplus, il n'est pas toujours compris entre 0 et N.

Que $\hat{n}$ puisse être fractionnaire n'est pas spécialement gênant, surtout dans le cas où il est déduit d'une moyenne telle que $\bar{X}_Q$. De façon analogue, personne ne se soucie du fait que l'indice synthétique de fécondité (le nombre moyen d'enfants par femme) ne soit pas un entier.

Toutefois, d'aucuns sont inconfortables à l'idée que $\hat{n}$ puisse être négatif ou supérieur à N. On peut facilement corriger ces défauts en posant par exemple

$$\check{n} = \min\{N, \max(0, \lfloor \hat{n} \rfloor)\},$$

où $\lfloor \hat{n} \rfloor$ dénote la partie entière de $\hat{n}$. L'estimateur $\check{n}$ résultant est alors plus facile à interpréter que $\hat{n}$, mais il n'est plus sans biais. Pour le vérifier, exprimons $X = V_n + W_{N-n}$ comme une somme de variables aléatoires binomiales indépendantes, à savoir

$$V_n = \sum_{i=1}^{n} X_i \sim \mathcal{BIN}[n, (1+p)/2], \quad W_{N-n} = \sum_{i=n+1}^{N} X_i \sim \mathcal{BIN}[N-n, (1-p)/2].$$

Il s'ensuit que, pour tout entier $x \in \{0, \ldots, N\}$,

$$\Pr(X = x) = \sum_{y=\max(0,x+n-N)}^{\min(x,n)} \Pr(V_n = y) \Pr(W_{N-n} = x - y), \tag{3.1}$$

puisque l'on a forcément $0 \leq y \leq x$ et $0 \leq x - y \leq N - n$.

Les probabilités données dans la formule (3.1) se calculent de façon explicite mais leur expression est complexe. Une importante simplification se produit toutefois lorsque $n = 0$, car alors $X = W_N$

est une variable binomiale de paramètres N et $(1-p)/2$. La probabilité d'avoir $X > N(1-p)/2$ étant non nulle, on peut avoir $\hat{n} > 0$ et donc $\check{n} > 0$. Par conséquent, $\mathrm{E}(\check{n}) > 0 = n$, démontrant ainsi que l'estimateur $\check{n}$ est bel et bien biaisé.

On peut aussi exploiter la formule (3.1) pour déterminer l'estimateur à vraisemblance maximale du nombre de personnes vaccinées. Par définition, il s'agit de l'entier $n \in \{0, \dots, N\}$ qui maximise la fonction

$$\begin{aligned} L(n) &= \sum_{y=\max(0,x+n-N)}^{\min(x,n)} \binom{n}{y} \left(\frac{1+p}{2}\right)^{y} \left(\frac{1-p}{2}\right)^{n-y} \binom{N-n}{x-y} \left(\frac{1-p}{2}\right)^{x-y} \left(\frac{1+p}{2}\right)^{N-n-x+y} \\ &= \sum_{y=\max(0,x+n-N)}^{\min(x,n)} \binom{n}{y}\binom{N-n}{x-y} \left(\frac{1-p}{2}\right)^{x+n-2y} \left(\frac{1+p}{2}\right)^{N-(x+n-2y)}. \end{aligned}$$

Des résultats classiques de statistique mathématique permettent alors de conclure que la solution est convergente et donc asymptotiquement sans biais. Toutefois, elle n'est pas explicite, ce qui la rend moins attrayante dans un cadre pédagogique. Pour cette raison, nous nous limiterons par la suite à l'estimateur des moments, soit $\hat{n}$.

4 Calcul de la marge d'erreur

Alors que la valeur de n est constante, bien qu'inconnue, celle de $\hat{n}$ dépend des tirages aléatoires effectués indépendamment par les N membres du groupe. Il s'agit d'une variable aléatoire dont la loi est tributaire de celle de X précisée en (3.1).

Une mesure de la variabilité de $\hat{n}$ est donnée par sa variance, laquelle dépend de celle de X. Or, $\mathrm{var}(X) = \mathrm{var}(V_n) + \mathrm{var}(W_{N-n})$ puisque les variables aléatoires V_n et W_{N-n} sont indépendantes. De plus, on a $1 - (1+p)/2 = (1-p)/2$, d'où il découle que

$$\mathrm{var}(V_n) = n\left(\frac{1+p}{2}\right)\left(\frac{1-p}{2}\right), \quad \mathrm{var}(W_{N-n}) = (N-n)\left(\frac{1+p}{2}\right)\left(\frac{1-p}{2}\right)$$

et donc

$$\mathrm{var}(X) = N\left(\frac{1+p}{2}\right)\left(\frac{1-p}{2}\right).$$

Ainsi la variance de X est-elle la même que si cette variable était binomiale de paramètres N et $(1+p)/2$ ou $(1-p)/2$. On déduit en outre de la formule (2.1) que

$$\mathrm{var}(\hat{n}) = \frac{1}{p^2}\,\mathrm{var}(X) = N\sigma_p^2, \tag{4.1}$$

où

$$\sigma_p^2 = \left(\frac{1+p}{2p}\right)\left(\frac{1-p}{2p}\right).$$

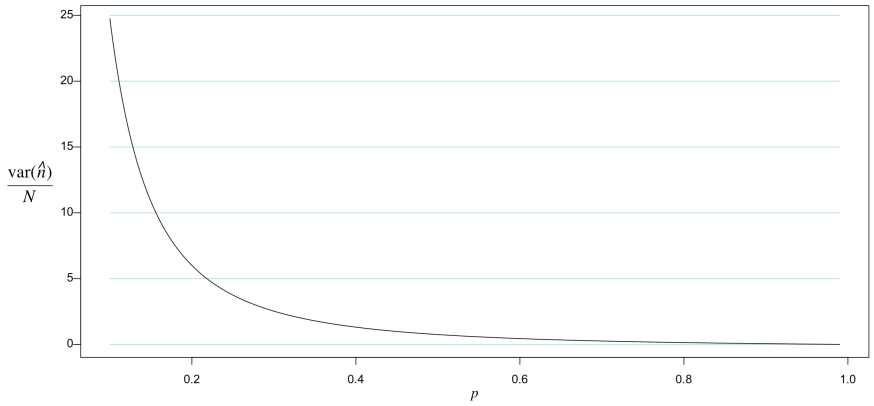

FIGURE 1 – Graphe du rapport $\mathrm{var}(\hat{n})/N$ déduit de (4.1) en fonction de $p \in]0,1[$.

On remarque que cette expression ne dépend pas du paramètre d'intérêt, à savoir n, mais qu'elle n'est fonction que de deux constantes connues, soit la taille N du groupe et le paramètre p.

Voici deux questions qui surgissent immédiatement et que l'on pourrait discuter en classe :

a) Pour quelle valeur de p la variance est-elle minimale ?
b) Cette valeur est-elle réalisable ? Sinon, quelle valeur de p constituerait un bon compromis ?

Comme l'illustre la figure 1, l'application $p \mapsto (1-p^2)/(2p)^2$ est décroissante sur tout l'intervalle $]0,1]$. La variance de l'estimateur $\hat{n}$ serait minimisée en $p = 1$, auquel cas $\mathrm{var}(\hat{n}) = 0$ car il n'y aurait alors plus de composante aléatoire au sondage : dans ce cas limite, tous les répondants dévoileraient leur statut vaccinal sans mentir et on aurait tout simplement $\hat{n} = X = n$.

La décroissance de $\mathrm{var}(\hat{n})$ en fonction de p fait ressortir le dilemme qui existe entre quête d'information et protection de la vie privée. Plus p se rapproche de 1, plus l'estimation sera précise mais moins les membres du groupe se sentiront à l'aise de participer au sondage et de répondre honnêtement à la question posée.

Idéalement, le choix de p pourrait faire l'objet d'une décision de groupe. La figure 1 pourrait aider les membres à atteindre un juste équilibre et à maximiser les chances de succès de l'opération. On y voit que la décroissance du rapport $\mathrm{var}(\hat{n})/N$ déduit de (4.1) est de plus en plus lente à mesure que p augmente. Il semble raisonnable de prendre au moins $p \geq 1/2$.

Pour éclairer davantage le choix de p, on peut déterminer à l'avance une marge d'erreur approximative pour $\hat{n}$ en faisant appel au théorème central limite. En effet, la proportion de

participants ayant répondu « oui » peut s'exprimer comme suit

$$\frac{\hat{n}}{N} = \frac{\bar{X}_N - (1-p)/2}{p}$$

en terme de la moyenne $\bar{X}_N = (X_1 + \cdots + X_N)/N$ de N variables aléatoires de Bernoulli mutuellement indépendantes. Or, bien que ces variables ne soient pas identiquement distribuées, un résultat du mathématicien russe Alexandre Liapounov (1857–1918) permet d'affirmer que la loi limite de $\bar{X}_N$ est néanmoins gaussienne (voir l'Annexe 1 pour de plus amples détails).

Quoiqu'en pratique la taille N du groupe soit évidemment finie, l'approximation gaussienne paraît adéquate, comme en fait foi la figure 2 (page 26), qui illustre la fonction de probabilité de la variable X lorsque $p = 1/2$ et $n/N = 0{,}75$ (gauche) ou $n/N = 0{,}95$ (droite) pour $N \in \{20, 40, 60, 80\}$. Bien que la loi soit légèrement asymétrique, l'approximation gaussienne semble raisonnable. Pour des résultats plus fins, voir par exemple [3], [6] ou [10].

Sachant que

$$\hat{n} - n \approx \mathcal{N}(0, N\sigma_p^2),$$

on sait que $\Pr(|\hat{n} - n|/\sqrt{N\sigma_p^2} \leq 2) \approx 0{,}95$, de sorte que les bornes d'un intervalle de confiance asymptotique de niveau 95% pour n sont données par la formule

$$\hat{n} \pm 2\sigma_p\sqrt{N}, \tag{4.2}$$

dans laquelle le facteur multiplicatif 2 qui précède la racine carrée est en fait une valeur approchée du 97,5$^{\text{e}}$ quantile de la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$, soit 1,959964…

De façon semblable, les bornes d'un intervalle de confiance asymptotique à 95% pour la proportion n/N d'individus vaccinés sont données par la formule

$$\frac{\hat{n}}{N} \pm 2\frac{\sigma_p}{\sqrt{N}}. \tag{4.3}$$

Quoiqu'elles soient imparfaites, les marges d'erreur approximatives données en (4.2) et en (4.3) ont l'avantage d'être très faciles d'utilisation. Elles permettent en outre de déterminer par avance la longueur des intervalles de confiance déduits de ces formules. Ainsi est-il possible de choisir p en fonction du degré de précision voulu, dans le respect de la vie privée.

À titre d'illustration, la marge d'erreur sur l'estimation de n vaut $\sqrt{3N} \approx 1{,}73\sqrt{N}$ quand $p = 1/2$; elle est réduite à $1{,}33\sqrt{N}$ si $p = 3/5$ et n'est plus que de $0{,}88\sqrt{N}$ lorsque $p = 3/4$.

Le nomogramme de la figure 3 (page 27) montre aussi (en noir) le comportement de l'estimateur $\hat{n}/N$ de la couverture vaccinale quand $p = 3/4$. Il est clair que cette estimation ne dépend pas de la taille $N \in \{20, 40, 60, 80\}$ du groupe mais seulement de la proportion n/N réelle

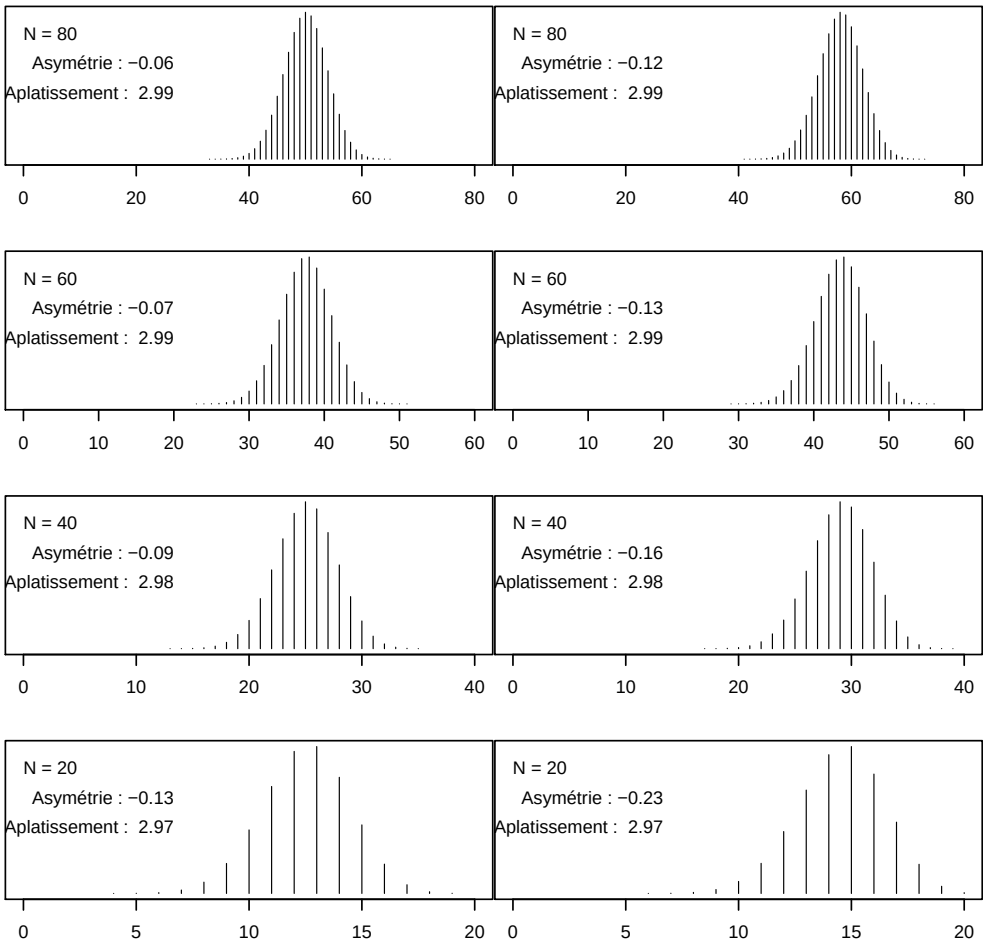

FIGURE 2 – Représentation de la loi de X lorsque $p = 1/2$ et que $n/N = 0{,}75$ (gauche) ou $n/N = 0{,}95$ (droite) pour $N \in \{20, 40, 60, 80\}$.

mais inconnue d'individus vaccinés au sein du groupe. En revanche, les limites de l'intervalle de confiance asymptotique de niveau 95%, représentées par des lignes bleues et rouges sur le nomogramme, se resserrent autour de l'estimation à mesure que N croît.

Une autre façon d'illustrer le dilemme entre précision et discrétion consiste à fixer d'abord l'entier k pour lequel $\hat{n} \pm k$ constitue un intervalle de confiance asymptotique de niveau 95% pour n et à déduire la valeur minimale de p pour répondre à cette exigence.

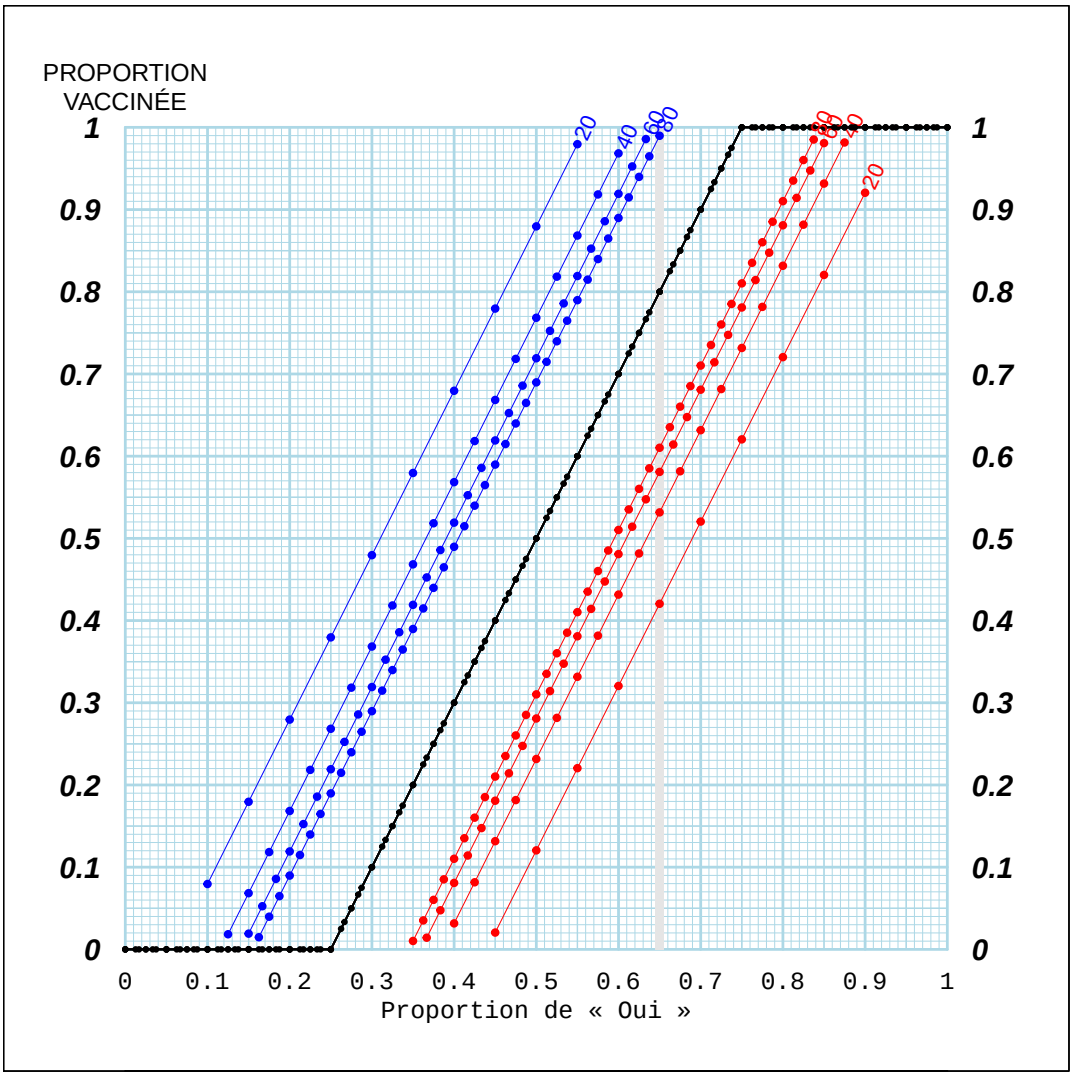

FIGURE 3 – Nomogramme permettant d'inférer la proportion de gens vaccinés (axe vertical) dans un groupe de taille $N \in \{20, 40, 60, 80\}$ à partir de la proportion de « oui » (X/N, axe horizontal) observés dans un sondage réalisé selon la MQA dans lequel la probabilité d'avoir à répondre à la question « Êtes-vous une personne (pleinement) vaccinée ? » est $p = 3/4$. La courbe en noir montre le comportement de l'estimateur $\hat{n}/N$ en fonction de X/N. Les bornes de l'intervalle de confiance asymptotique ponctuel de niveau 95% sont indiquées en rouge (limite inférieure) et en bleu (limite supérieure) pour différentes tailles de groupe.

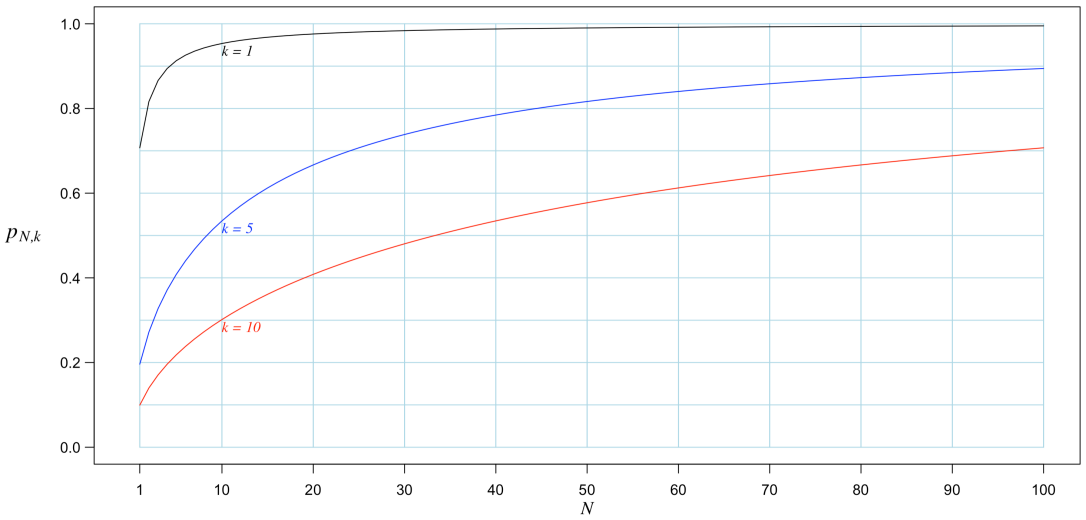

FIGURE 4 – Graphe de la probabilité $p_{N,k}$ définie en (4.4) en fonction de N lorsque $k = 1$ (noir), $k = 5$ (bleu) et $k = 10$ (rouge).

La taille N du groupe étant donnée, on doit choisir p tel que

$$2\sigma_p\sqrt{N} = k.$$

La seule solution de cette équation quadratique en p dans l'intervalle ouvert $]0, 1[$ est

$$p_{N,k} = \sqrt{\frac{N}{N + k^2}}\,. \tag{4.4}$$

La constante $p_{N,k}$ diminue lorsque k augmente, mais elle croît aussi avec N. Il faut donc augmenter p et ainsi sacrifier un certain degré de protection si

a) la taille N du groupe est fixée et on souhaite diminuer la marge d'erreur k ;
b) la marge d'erreur k est fixée et la taille N du groupe augmente.

Ces constatations sont illustrées à la figure 4, qui montre que $p_{N,k}$ tend vers 1 assez rapidement à mesure que N augmente, surtout lorsque $k = 1$ (courbe noire), de façon moins prononcée lorsque $k = 5$ (courbe bleue) et encore moins lorsque $k = 10$ (courbe rouge).

5 Autre façon d'améliorer la précision de l'estimation

La marge d'erreur relativement grande associée à la MQA est le prix à payer pour la protection des répondants. Cependant, si ceux-ci sont disposés à répéter le sondage, l'effet du « bruit » induit par la randomisation peut être réduit. Considérant l'intérêt que tous portent au sujet de la vaccination, et compte tenu du fait que le sondage peut être réalisé en ligne et facilement répété, disons Q fois, il y a fort à parier que bien des groupes voudront se prêter au jeu.

La répétition de l'enquête offre en outre une occasion unique d'observer l'effet du hasard sur le nombre total ou la proportion de « oui » d'une répétition à l'autre. Ces valeurs, dénotées $X_1, \ldots, X_Q$, sont mutuellement indépendantes même si elles s'appuient sur les mêmes répondants, puisque c'est le mécanisme de randomisation qui leur confère un caractère aléatoire.

Chacune des valeurs $X_1, \ldots, X_Q$ conduit à une estimation distincte $\hat{n}_1, \ldots, \hat{n}_Q$ du nombre d'individus vaccinés dans le groupe. Soit alors

$$\bar{n} = (\hat{n}_1 + \cdots + \hat{n}_Q)/Q$$

la valeur moyenne du nombre de personnes vaccinées prise sur l'ensemble des Q répétitions. Il découle de l'indépendance mutuelle des estimations $\hat{n}_1, \ldots, \hat{n}_Q$ que

$$\mathrm{var}(\bar{n}) = \sigma_p^2 N/Q.$$

Si par exemple les participants sont disposés à participer au sondage $Q = 4$ fois de suite, la marge d'erreur s'en trouvera réduite de moitié. De façon équivalente, les bandes de confiance (ponctuelles) associées à $N = 80$ dans le nomogramme de la figure 3 sont celles qui correspondent à une estimation de la proportion $\bar{n}/N$ d'individus vaccinés sur la base de quatre sondages indépendants réalisés sur le même groupe de $N = 20$ personnes.

En effectuant le sondage à Q reprises plutôt qu'une seule fois, on peut donc améliorer la précision de l'estimation sans changer la valeur de p, c'est-à-dire sans affecter le degré de protection des participants. L'interaction entre les paramètres p et Q déterminant la marge d'erreur $\pm k$ d'un intervalle de confiance de niveau 95% asymptotique est dictée par la relation

$$2\sigma_p \sqrt{N/Q} = k. \tag{5.1}$$

Bien entendu, on pourrait aussi opter pour un autre niveau de confiance que 95%, disons $100 \times (1 - \alpha)\%$, où $\alpha \in (0, 1)$. Le facteur 2 dans l'équation (5.1) serait alors remplacé partout par $z_{\alpha/2}$, où z_α dénote le quantile supérieur d'ordre α de la loi normale centrée réduite $\mathcal{N}(0, 1)$.

6 Conclusion

En décrivant la manière dont un mécanisme de réponse aléatoire simple peut être présenté et utilisé en classe pour aborder une question d'actualité mais sensible comme celle de la vaccination contre la COVID-19, nous espérons avoir montré la pertinence de cette technique et peut-être aussi avoir convaincu certains lecteurs qu'ils pourraient se servir de la MQA en classe afin d'illustrer divers concepts statistiques de base qui interviennent dans sa mise en œuvre.

Dans cet article, tous les calculs ont été effectués en supposant que la question auxiliaire à laquelle répondent les participants concerne le jet d'une pièce de monnaie équilibrée. Ce choix a été fait à des fins pédagogiques et pour assurer une forme de protection légale au sondeur. En effet, celui-ci ne peut pas être accusé d'avoir demandé à une personne son statut vaccinal puisqu'il n'est même pas certain que cette dernière ait eu à envisager la question.

ll existe toutefois de très nombreuses variantes de la MQA et chacune conduit à une analyse statistique différente. Par exemple, comme dans l'article de Warner [14], on pourrait demander à chaque participant de choisir aléatoirement entre les énoncés « Je suis vacciné » et « Je ne suis pas vacciné » et de simplement répondre par oui ou non à l'énoncé sélectionné.

La version de la MQA employée ici s'inspire des variantes de Horvitz et coll. [9] et de Greenberg et coll. [8]. Une généralisation immédiate consisterait à remplacer la pièce équilibrée par une pièce dans laquelle la probabilité de pile est un nombre $r \in]0, 1[$ connu. Moors [11] a montré que cette approche peut accroître la précision. Toutefois, l'analyse statistique est plus complexe car la variance de l'estimateur dépend alors du nombre n de personnes vaccinées.

D'autres options sont recensées par Blair et coll. [2], qui décrivent les stratégies d'estimation adaptées aux différents scénarios et qui en comparent l'efficacité. Au-delà de l'approche fréquentiste classique employée ici et à laquelle sont associés les estimateurs des moments et à vraisemblance maximale, on peut envisager une approche bayésienne, comme le fait par exemple O'Hagan [12], qui présente des estimateurs de Bayes linéaires.

Les lecteurs férus d'histoire prendront plaisir à lire l'article de Fienberg et Jazairi [5] retraçant les contributions de Warner à la technologie de l'information statistiquement équilibrée. Dans un tout autre registre, on pourra aussi consulter sur le site de la revue *Accromath* un court article de vulgarisation sur la MQA accessible aux élèves du secondaire [7].

Dans un contexte de pandémie, la vaccination est généralement perçue comme le meilleur moyen de protection qui soit. En même temps, chacun est libre de ses choix et le statut vaccinal est considéré comme une affaire privée. Tous ces éléments contribuent à ce que les enseignant.e.s et les étudiant.e.s soient intéressés par le résultat d'un sondage réalisé en classe sur cette question.

En somme, l'occasion est belle de mettre en pratique la méthode de la question aléatoire et de se familiariser du même coup avec un certain nombre de concepts statistiques. Nous espérons

que cet article stimulera cette réflexion.

À titre d'exercice récapitulatif susceptible d'engendrer un intéressant débat en classe, on pourra se demander à partir de quel nombre Q de répétitions on serait convaincu hors de tout doute raisonnable que dans le cas $N = 1$, l'unique répondant est ou n'est pas vacciné.

Annexe 1 : Démonstration de l'approximation gaussienne

Soit $X_1, X_2, \ldots$ une suite de variables aléatoires de Bernoulli mutuellement indépendantes. À moins que ces variables soient toutes de même loi, on dit qu'il s'agit d'essais de Poisson [4]. Pour tout entier $i \in \{1, 2, \ldots\}$, posons $\mathrm{E}(X_i) = p_i \in]0, 1[$, $Y_i = X_i - p_i$ et $\sigma_i^2 = p_i(1 - p_i)$. Pour tout entier $m \in \{1, 2, \ldots\}$, soient en outre $S_m = Y_1 + \cdots + Y_m$ et $s_m^2 = \sigma_1^2 + \cdots + \sigma_m^2$. Le théorème de Liapounov stipule que la suite S_m/s_m converge en loi vers une variable aléatoire gaussienne centrée réduite quand $m \to \infty$ pourvu qu'il existe un nombre $\delta > 0$ tel que

$$\lim_{m\to\infty} \frac{1}{s_m^{2+\delta}} \sum_{i=1}^{m} \mathrm{E}|Y_i|^{2+\delta} = 0.$$

Pour vérifier cette condition, posons $\delta = 1$ et observons que pour tout entier $i \in \{1, \ldots, m\}$,

$$\mathrm{E}|Y_i|^3 = p_i(1 - p_i)^3 + (1 - p_i)p_i^3 = \sigma_i^2\{(1 - p_i)^2 + p_i^2\} \leq \sigma_i^2.$$

Il s'ensuit que $\mathrm{E}|Y_1|^3 + \cdots + \mathrm{E}|Y_m|^3 \leq s_m^2$ et donc que la condition de Liapounov est remplie dès que $s_m^2/s_m^3 \to 0$ quand $m \to \infty$, c'est-à-dire dès que la suite s_m diverge. Or cette condition est satisfaite dans le cas traité dans le corps de l'article puisque $\Pr(X_i) = (1 - p)/2$ ou $\Pr(X_i) = (1 + p)/2$ et donc $p_i(1 - p_i) = (1 - p^2)/4$ pour tout entier $i \in \{1, 2, \ldots\}$, de sorte que

$$s_m = \sqrt{m(1 - p^2)/4},$$

et donc $s_m \to \infty$ quand $m \to \infty$, ce qui permet de conclure.

Annexe 2 : Comment réaliser un sondage en ligne

Voici comment réaliser un sondage électronique afin d'estimer instantanément, de façon anonyme et dans le respect de la vie privée, le nombre de personnes vaccinées au sein d'un groupe.

Sachez d'abord que les options sont nombreuses, comme vous le constaterez en cherchant « sondages en ligne » sur Internet. Si certains sites commerciaux sont fiables et sophistiqués, d'autres sont éphémères ou suspects.

L'outil illustré ici est gratuit pour les sondages d'au plus 100 participants et ses fonctionnalités sont suffisantes pour nos fins. Certains logiciels d'enseignement peuvent également faire l'affaire.

Procédure

Le site `http://www.vevox.com` se présente comme le « meilleur outil de sondage et de questions/réponses » disponible sur Internet.

Dans le coin supérieur droit de la page d'accueil, on peut créer gratuitement un compte en précisant son adresse courriel et en choisissant un mot de passe à 8 caractères.

On doit fournir un nom, un prénom, une raison sociale, un titre d'emploi, un numéro de téléphone et un pays de résidence. On doit aussi proposer une adresse unique (URL) pour le tableau de bord. Si on entre le sigle de cours math101, par exemple, le tableau de bord s'appellera `math101.vevox.com` et cet identifiant sera partagé lors de la tenue du sondage.

Une fois le compte créé, on est invité à visionner une vidéo de bienvenue ou à démarrer une session. La vidéo vaut la peine d'être vue au moins une fois, mais il faut en interrompre le visionnement souvent car les instructions défilent rapidement. Elles sont résumées ci-dessous.

1) Démarrer une nouvelle session et la nommer.

   Vevox valide l'adresse par courriel lors de la première utilisation.

2) Créer le sondage (on peut faire fi de « Get started »).

3) Choisir la modalité « question à choix multiples » et formuler une question.

   Suggestion de formulation : « Faites appel au mécanisme convenu pour choisir aléatoirement l'une des deux questions suivantes et répondez-y honnêtement :
   (A) Êtes-vous une personne (pleinement) vaccinée ? ou (B) Avez-vous obtenu pile ? »

4) Entrer oui/non comme choix de réponse.

5) Appuyer sur le bouton « Add » pour sauvegarder le sondage.

Pour lancer la consultation, partager son écran avec l'auditoire après avoir appuyé sur « Present » (coin supérieur droit), ce qui a pour effet d'afficher les instructions à suivre pour participer au sondage (y compris, si on le souhaite, par le biais d'un code QR).

Utiliser le panneau de configuration au bas de l'écran pour démarrer la collecte de données. Une fois qu'elle est complétée, mettre fin au sondage. Les résultats s'affichent alors automatiquement à l'écran et sur les appareils des participants.

Pour recommencer, appuyer d'abord sur les 3 points verticaux dans le coin inférieur droit et remettre le compteur à zéro (chaque question est traitée comme un sondage distinct).

Le sondage peut être consulté dans le tableau de bord ou présenté en mode plein écran.

Références

[1] Bellhouse, David R. *Estimation de la corrélation dans les plans à réponses randomisées*. Techniques d'enquête, Vol. 21 (1995), no 1, pp. 15–21.

[2] Blair, Graeme, Imai, Kosuke & Zhou, Yang-Yang. *Design and analysis of the randomized response technique*. Journal of the American Statistical Association, Vol. 110 (2015) no 511, pp. 1304–1318.

[3] Butler, Ken & Stephens, Michael A. *The distribution of a sum of independent binomial random variables*. Methodology and Computing in Applied Probability, Vol. 19 (2017), no 2, pp. 557–571.

[4] Feller, William. *An Introduction to Probability Theory and Its Applications*, 3e édition, Vol. 1. Wiley, New York, 1968.

[5] Fienberg, Steven E. & Jazairi, Nuri. *Contributions de Stanley Warner à la technologie de l'information statistiquement équilibrée*. Techniques d'enquête, Vol. 21 (1995), no 1, pp. 7–13.

[6] Frey, Jesse & Pérez, Andrés. *Exact binomial confidence intervals for randomized response*. The American Statistician, Vol. 66 (2012), no 1, pp. 8–15.

[7] Genest, Christian. *La méthode de Warner*. Accromath, Vol. 15 (2020), no 1, pp. 4–5.

[8] Greenberg, Bernard G., Abul-Ela, Abdel-Latif A., Simmons, Walt R. & Horvitz, Daniel G. *The unrelated question randomized response model : Theoretical framework*. Journal of the American Statistical Association, Vol. 64 (1969), no 326, pp. 520–539.

[9] Horvitz, Daniel G., Shah, Babubhai V. & Simmons, Walt R. *The unrelated question randomized response model*. Social Statistics Section Proceedings of the American Statistical Association, 1967, pp. 65–72.

[10] Liu, Boxiang & Quertermous, Thomas. *Approximating the sum of independent non-identical binomial random variables*. The R Journal, Vol. 10 (2018), no 1, pp. 472–483.

[11] Moors, J. J. A. *Optimization of the unrelated question randomized response model*. Journal of the American Statistical Association, Vol. 66 (1971), no 335, pp. 627–629.

[12] O'Hagan, Anthony. *Bayes linear estimators for randomized response models*. Journal of the American Statistical Association, Vol. 82 (1987), no 398, pp. 580–585.

[13] Särndal, Carl-Erik. *Stanley L. Warner : 1928–1992.* Techniques d'enquête, Vol. 21 (1995), no 1, pp. 3–4.

[14] Warner, Stanley L. *Randomized response : A survey technique for eliminating evasive answer bias*. Journal of the American Statistical Association, Vol. 60 (1965), no 309, pp. 63–69.