Chapitre 15 – Échantillonnage – Activité unique (corrigé)

Première partie

Étude de quelques définitions, pour comprendre le lexique de l'échantillonnage

Définitions. (*Source : Wikipedia*)

1. L'*échantillonnage* [...] désigne les méthodes de sélection d'un sous-ensemble d'individus (un échantillon) à l'intérieur d'une population pour estimer les caractéristiques de l'ensemble de la population.
2. Un *échantillon* est un ensemble d'individus représentatifs d'une population.

Question 1. Selon vous, quels pourraient être les avantages de l'échantillonnage lorsque l'on réalise un sondage ?

Proposition de réponse 1. Lorsque l'on réalise un sondage, il se peut que la taille de la population soit relativement élevée. Par exemple, dans le cadre d'un sondage politique, la taille de la population est celle de l'ensemble des personnes ayant le droit de vote (en France, il y avait environ $49,3$ millions d'électeurs en mai 2024). Interroger les $49,3$ millions d'électeurs serait chronophage, et les résultats seraient inexploitables ; le temps du vote, certains électeurs auraient pu changer d'avis, par exemple. C'est en cela que l'échantillonnage est bénéfique : il permet d'éviter d'avoir à interroger toute la population, mais seulement un échantillon qui reflètera une approximation de l'avis de la population totale.

Question 2. Plaçons-nous dans le cadre d'un sondage qui aurait lieu à la veille d'une élection politique d'une grande importance. Quels critères proposeriez-vous pour choisir des individus représentatifs de la population française ?

Proposition de réponse 2. Voici une proposition de quelques critères qui pourraient permettre de constituer l'échantillon le plus représentatif possible :

1. La distinction de genre,
2. La distinction suivant l'âge,
3. La distinction suivant la catégorie socio-professionnelle,
4. La distinction géographique (rurale, péri-urbaine, urbaine),
5. La distinction suivant l'appartenance à un parti politique,
6. La distinction suivant la certitude d'aller voter ou non...

Deuxième partie

Quelques mises en situation pratiques : sondages, résultats et interprétations

Situation 1. En dessous de cette publicité, il est indiqué *Scorage clinique, 53 femmes.*

Question 1. Que penser du choix de l'échantillon ?

Proposition de réponse 1. L'échantillon a été testé sur 53 femmes. Partant du principe que le produit n'est lancé qu'en France, et que sa cible serait constituée uniquement des femmes majeure, cela ferait environ 29 millions d'utilisatrices potentielles. Raisonnablement, on ne peut pas affirmer qu'un échantillon de 53 femmes soit suffisant, surtout si l'on pense au rapport entre les bénéfices et les risques des produits cosmétiques.

Point culture. L'autorité britannique de supervision de la publicité a jugé trompeuse une publicité magazine pour de la crème anti-âge d'une certaine marque, et interdit sa diffusion sous sa forme actuelle. La publicité en question, parue dans des magazines britanniques, affirmait que les utilisatrices de la crème de jour pouvaient « affronter l'avenir avec une peau plus ferme », mais sans expliquer clairement que la crème n'avait pas d'effet visible permanent. Par ailleurs, la publicité affirmait que « 126 femmes étaient d'accord » avec cette affirmation, mais sans indiquer combien, au total, avaient essayé la crème anti-âge, ce qui ne permettait pas aux lectrices de se faire une précise de son efficacité.

Situation 2. (*Source : Wikipedia*) Emmanuel Macron ayant été élu pour un second mandat au second tour de l'élection présidentielle de 2022, la Constitution lui interdit de renouveler son mandat une deuxième fois consécutive. La question de la succession au président de la République se pose alors au sein de la majorité présidentielle, ce qui amène les sondeurs à conduire des enquêtes d'opinions sur le sujet. Le tableau ci-dessous ne comprend que les sondages les plus récents pour chaque institut.

Sondeur	Date	Échantillon		 Attal (REN)	 Bayrou (MoDem)	 Darmanin (REN)	 Le Maire (REN)	 Philippe (HOR)
Ifop [archive]	3-4 avril 2024	1 028	Ensemble des Français	**63 %**	31 %	49 %	50 %	55 %
			Partisans de la majorité	**92 %**	54 %	68 %	75 %	87 %
Odoxa [archive]	14-15 février 2024	1 005	Ensemble des Français	44 %	19 %	23 %	30 %	**49 %**
			Partisans de la majorité	82 %	23 %	39 %	70 %	**86 %**

Question 2. D'un sondage à l'autre, les résultats sont tout de même assez différents. Comment expliquez-vous ces différences ?

Proposition de réponse 2. Les tailles des échantillons sont assez comparables (1 028 pour le premier sondage, 1 005 pour le second), ce qui ne permet pas d'expliquer de tels écarts. En revanche, il y a d'autres raisons qui permettent d'expliquer ces écarts, dont voici quelques exemples :

1. Les résultats entre les lignes *Ensemble des Français* et *Partisans de la majorité* sont évidemment différents, pour la simple raison que les partisans d'un parti politique ont plutôt vocation à voter pour leur propre parti, ou un parti très apparenté.
2. Les sondages ont été réalisés à des dates différentes. Aussi, il est possible qu'un évènement qui soit arrivé entre-temps puisse changer les convictions des électeurs.
3. Les sondages ont été réalisés par des sondeurs différents. Aussi, il est possible qu'un institut de sondage (privé) choisisse davantage les personnes sondées selon ses propres critères, pour favoriser le choix d'un candidat parmi un autre.

Situation 3. Un site de commerce en ligne a lancé une enquête auprès de 300 clients, choisis au hasard parmi ceux ayant effectué des achats sur leur site. Le temps de livraison a été jugé « satisfaisant » par 160 des personnes interrogées.

Question 3. Calculer la fréquence f de clients ayant jugé « satisfaisant » le temps de livraison dans cet échantillon. Donner un intervalle de confiance au seuil de 95% de la proportion p de clients de ce site Internet satisfaits par le temps de livraison.

Proposition de réponse 3. La fréquence de clients ayant jugé « satisfaisant » le temps de livraison dans cet échantillon est $f = \frac{160}{300} \approx 0,533$; soit environ $53,3\%$. L'intervalle de confiance au seuil de 95% de la proportion p de clients de ce site Internet satisfaits par le temps de livraison est alors de :

$$\left[f - \frac{1}{\sqrt{n}} \,;\, f + \frac{1}{\sqrt{n}}\right] = \left[0,533 - \frac{1}{\sqrt{300}} \,;\, 0,533 + \frac{1}{\sqrt{300}}\right] = [0,475 \,;\, 0,591]$$

où les bornes ont été arrondies 10^{-3} près.

Situation 4. En 1976, dans un comté du Texas, Rodrigo Partida était condamné à huit ans de prison, après une suspicion de cambriolage dans une résidence privée du Texas. Il attaqua ce jugement, au motif que la désignation des jurés de ce comté était discriminante à l'égard des américains d'origine mexicaine. En effet, 79% de la population de ce comté est d'origine mexicaine, et sur les 870 personnes convoquées pour être jurés, seules 339 d'entre elles étaient d'origine mexicaine.

Question 4. Peut-on dire que la constitution des jurys est faite de façon aléatoire ?

Proposition de réponse 4. Ici, on s'intéresse à la représentativité des américains d'origine mexicaine. Dans l'échantillon, c'est-à-dire dans les jurés du comté ayant participé au jugement, la fréquence des américains d'origine mexicaine est $f = \frac{339}{870} \approx 0,3897$; soit environ $38,97\%$. L'intervalle de confiance au seuil de 95% de la proportion p d'américains d'origine mexicaine dans tout le comté serait alors de :

$$\left[f - \frac{1}{\sqrt{n}} \,;\, f + \frac{1}{\sqrt{n}}\right] = \left[0,3897 - \frac{1}{\sqrt{870}} \,;\, 0,3897 + \frac{1}{\sqrt{870}}\right] = [0,3558 \,;\, 0,4236]$$

où les bornes ont été arrondies 10^{-4} près. Or, on sait que 79% de la population de ce comté est d'origine mexicaine, et $0,79 \notin [0,3558 \,;\, 0,4236]$. **La constitution des jurys est donc douteuse.**

Troisième partie
Retour aux mathématiques

Pour un intervalle I donné, on définit sa *longueur* $\ell(I)$ comme l'écart entre ses deux bornes, peu importe l'orientation des crochets. Si l'une des bornes est infinie, on conviendra que sa longueur est égale à $+\infty$. Par exemple, la longueur de l'intervalle $I =]2 \,;\, 7]$ est égale à $\ell(I) = 7 - 2 = 5$.

Question 1. Déterminer la longueur des intervalles $I = [0 \,;\, 1]$, $J =]-2 \,;\, 1]$ et $K = [0 \,;\, +\infty[$.

Proposition de réponse 1. La longueur de l'intervalle K est égale à $+\infty$, puisque l'une des bornes de K est elle-même infinie. Sinon, $\ell(I) = 1 - 0 = 1$ et $\ell(J) = 1 - (-2) = 3$.

On rappelle que si un sondage produit une estimation f pour la présence d'un caractère dans un échantillon, alors il y a (au moins) 95% de chances que la proportion p de présence d'un caractère dans la population totale appartienne à l'intervalle de confiance :

$$\left[f - \frac{1}{\sqrt{n}} \,;\, f + \frac{1}{\sqrt{n}}\right]$$

Question 2. Déterminer la longueur de l'intervalle de confiance.

Proposition de réponse 2. $\ell = \left(f + \frac{1}{\sqrt{n}}\right) - \left(f - \frac{1}{\sqrt{n}}\right) = f + \frac{1}{\sqrt{n}} - f + \frac{1}{\sqrt{n}} = \frac{2}{\sqrt{n}}$.

Question 3. Commenter la phrase suivante : « Quand le nombre n de personnes sondées augmente, alors l'intervalle de confiance se resserre autour de f ». Que peut-on en déduire lorsqu'un sondage est réalisé ?

Proposition de réponse 3. On voit effectivement que, lorsque la valeur de n augmente, la longueur de l'intervalle de confiance devient de plus en plus petite, donc resserrée autour de sa valeur centrale :

1. Le centre de l'intervalle de confiance est bien f, puisque la moyenne de ses bornes est :
$$\frac{\left(f-\frac{1}{\sqrt{n}}\right)+\left(f+\frac{1}{\sqrt{n}}\right)}{2}=\frac{f-\frac{1}{\sqrt{n}}+f+\frac{1}{\sqrt{n}}}{2}=\frac{2f}{2}=f$$
2. Un calcul rapide donne l'intuition que la longueur de l'intervalle devient de plus en plus petite lorsque n devient de plus en plus grand :
$$\frac{2}{\sqrt{5}}\approx 0,8944\,;\ \frac{2}{\sqrt{10}}\approx 0,6325\,;\ \frac{2}{\sqrt{100}}=0,2\,;\ \frac{2}{\sqrt{1\,000}}\approx 0,0632$$

On souhaiterait « contrôler l'erreur d'estimation », c'est-à-dire faire en sorte que l'intervalle de confiance soit le plus resserré possible. Imaginons par exemple que l'on souhaite avoir une erreur qui soit au maximum égale à 10%. Alors il faudrait que $\frac{2}{\sqrt{n}}\leq 0,1$, c'est-à-dire que $2\leq 0,1\sqrt{n}$, autrement dit $\sqrt{n}\geq\frac{2}{0,1}$, ce qui donne $\sqrt{n}\geq 20$, ou encore $n\geq 400$. Il faudrait interroger 400 personnes pour que l'erreur soit au maximum égale à 10%.

Question 4. En vous inspirant de l'exemple ci-dessus, résoudre une résolution pour déterminer quel serait le nombre minimum de personnes à interroger pour avoir :

1. Une erreur maximale de 5% ?
2. Une erreur maximale de 2% ?
3. Une erreur maximale de 1% ?
4. Une erreur maximale de $0,5\%$?

Proposition de réponse 4. Sur le même modèle que l'exemple ci-dessus, on a :

1. Pour une erreur maximale de 5%, il faudrait que $\frac{2}{\sqrt{n}}\leq 0,05$, c'est-à-dire que $2\leq 0,05\sqrt{n}$, autrement dit $\sqrt{n}\geq\frac{2}{0,05}$, ce qui donne $\sqrt{n}\geq 40$, ou encore $n\geq 1\,600$.
Il faudrait interroger 1 600 personnes pour que l'erreur soit au maximum égale à 5%.
2. Pour une erreur maximale de 2%, il faudrait que $\frac{2}{\sqrt{n}}\leq 0,02$, c'est-à-dire que $2\leq 0,02\sqrt{n}$, autrement dit $\sqrt{n}\geq\frac{2}{0,02}$, ce qui donne $\sqrt{n}\geq 100$, ou encore $n\geq 10\,000$.
Il faudrait interroger 10000 personnes pour que l'erreur soit au maximum égale à 2%.
3. Pour une erreur maximale de 1%, il faudrait que $\frac{2}{\sqrt{n}}\leq 0,01$, c'est-à-dire que $2\leq 0,01\sqrt{n}$, autrement dit $\sqrt{n}\geq\frac{2}{0,01}$, ce qui donne $\sqrt{n}\geq 200$, ou encore $n\geq 40\,000$.
Il faudrait interroger 40000 personnes pour que l'erreur soit au maximum égale à 1%.
4. Pour une erreur maximale de $0,5\%$, il faudrait que $\frac{2}{\sqrt{n}}\leq 0,005$, c'est-à-dire que $2\leq 0,005\sqrt{n}$, autrement dit $\sqrt{n}\geq\frac{2}{0,005}$, ce qui donne $\sqrt{n}\geq 400$, ou encore $n\geq 160\,000$.
Il faudrait interroger 160 000 personnes pour que l'erreur soit au maximum égale à $0,5\%$.