Avant le référendum

Introduction : Un référendum se prépare. Pour simplifier les choses, on se place dans l'hypothèse ou personne ne votera blanc (ou plus exactement on comptabilise comme "NON" tout ce qui n'est pas "OUI").

Soit p la proportion des gens qui voteront "OUI" au référendum. On effectue un sondage auprès d'un échantillon de n personnes prises au hasard. Parmi ces personnes, 60% répondent "OUI". On considère également que le nombre de personnes interrogées est très inférieur au nombre total de personnes concernées par le référendum.

Problème 1 *Que peut-on en déduire sur la proportion p ? Autrement dit, quelle est la fiabilité du sondage ? En quoi dépend-elle du nombre de personnes interrogées ?*

Au premier abord, on est tenté de répondre que, si l'on interroge par exemple 1000 personnes prises au hasard, l'échantillon est en gros représentatif, et donc que p est proche de $0,6$. Si l'on n'a interrogé que 10 personnes, il semble naturel de dire que l'échantillon est beaucoup moins représentatif, et que donc le sondage est moins fiable.

Cela dit, s'il y a disons un million de votants, il se peut très bien qu'au total seulement une personne sur quatre soit pour, et que notre échantillon de 1000 personnes peut être très particulier. C'est pourquoi il serait illusoire d'espérer affirmer quelquechose comme :

"Sachant que 600 personnes sur 1000 ont répondu "OUI" à notre sondage, le référendum SERA accepté."

En effet, les données disponibles, à savoir les réponses données par les personnes interrogées, sont plus que partielles : les opinions des électeurs non interrogés ne sont pas connues, un point c'est tout.

Bref, on ne peut rien affirmer de façon sûre, mais il semble raisonnable de dire que la situation décrite précédemment est peu probable. Dans les médias, les chiffres sont souvent donnés de façon brute, et l'on entend rarement parler de "fiabilité" d'un sondage, ou d'une "marge d'erreur de x%"... Voilà ce que nous allons essayer de formaliser ici.

Toutes les déductions sur la proportion réelle p que nous allons faire correspondent à ce que l'on appelle "inférence statistique" : ayant observé une réalisation, un tirage d'échantillon, on cherche à en tirer des conclusions sur la valeur réelle de p. L'inférence statistique est l'ensemble des méthodes permettant de formuler, en termes probabilistes, un jugement sur l'ensemble d'une population à partir d'observations faites sur un échantillon de cette population. Une introduction bien faite à ces problèmes peut être trouvée dans [3] ou [1].

1 Loi des grands nombres

En fait, lorsque l'on interroge n personnes "au hasard", c'est comme si l'on faisait n tirages simultanés d'une variable aléatoire X qui vaut 1 si la personne répond "OUI", ce avec une probabilité p, 0 si elle répond "NON", ce avec une probabilité $1-p$, l'inconnue étant la valeur de p, que l'on cherche à estimer.

1.1 Inégalité de Bienaymé-Chebychev

Une variable aléatoire X est un nombre qui est le résultat d'une expérience dont l'issue est incertaine. Par exemple, la valeur d'un lancé de dé, la durée de vie d'un noyau radioactif... Dans toute la suite, nous nous intéresserons à des variables aléatoires discrètes, c'est-à-dire ne pouvant prendre qu'un nombre au plus dénombrable de valeurs.

Théorème 1.1 *(Inégalité de Markov)*
Soit X *une variable aléatoire réelle positive. Soit* x_0 *un nombre positif. Alors*

$$x_0 \mathrm{P}(\mathrm{X} \geq x_0) \leq \mathrm{E}(\mathrm{X})$$

Démonstration Dans le cas d'une variable discrète, on a par définition :

$$\begin{aligned} \mathrm{E}(\mathrm{X}) &= \sum_{x \in \mathrm{E}} x \mathrm{P}(\mathrm{X} = x) \\ &= \sum_{x < x_0} x \mathrm{P}(\mathrm{X} = x) + \sum_{x \geq x_0} x \mathrm{P}(\mathrm{X} = x) \end{aligned}$$

Mais comme X est une variable aléatoire positive, le terme $\sum_{x<x_0} x \mathrm{P}(\mathrm{X} = x)$ est encore un réel positif. On en déduit que :

$$\begin{aligned} \mathrm{E}(\mathrm{X}) &\geq \sum_{x \geq x_0} x \mathrm{P}(\mathrm{X} = x) \\ &\geq x_0 \sum_{x \geq x_0} \mathrm{P}(\mathrm{X} = x) \end{aligned}$$

soit
$$\mathrm{E}(\mathrm{X}) \geq x_0 \mathrm{P}(\mathrm{X} \geq x_0)$$

□

N.B. Dans le cas d'une variable réelle quelconque, *i*.e. non nécessairement discrète, il suffit de remplacer les sommes par des intégrales, et l'on obtient immédiatement le même résultat.

Théorème 1.2 *(Inégalité de Bienaymé-Chebychev)*
Soit X *une variable aléatoire réelle et* m *son espérance. Alors pour tout* $a > 0$*, on a :*

$$\mathrm{P}(|\mathrm{X} - m| \geq a) \leq \frac{\mathrm{Var}(\mathrm{X})}{a^2}$$

Démonstration Il suffit d'appliquer l'inégalité de Markov à la variable aléatoire positive $(\mathrm{X}-m)^2$. On a :

$$a^2 \mathrm{P}\left((\mathrm{X}-m)^2 \geq a^2\right) \leq \mathrm{E}\left((\mathrm{X}-m)^2\right)$$

Mais $\mathrm{E}\left((\mathrm{X}-m)^2\right) = \mathrm{Var}(\mathrm{X})$ et la condition $(\mathrm{X}-m)^2 \geq a^2$ est équivalente à la condition $|\mathrm{X}-m| \geq a$, donc de même probabilité. On obtient comme annoncé :

$$\mathrm{P}(|\mathrm{X}-m| \geq a) \leq \frac{\mathrm{Var}(\mathrm{X})}{a^2}$$

□

1.2 Somme de variables aléatoires indépendantes

Si $\mathrm{X}_1, \ldots, \mathrm{X}_n$ sont des variables aléatoires discrètes, on dit qu'elles sont indépendantes si pour toutes valeurs $x_1, \ldots, x_n$, on a :

$$\mathrm{P}(\mathrm{X}_1 = x_1 \text{ et } \ldots \text{ et } \mathrm{X}_n = x_n) = \mathrm{P}(\mathrm{X}_1 = x_1) \times \ldots \times \mathrm{P}(\mathrm{X}_n = x_n)$$

Propriété 1.3 *Soient* $\mathrm{X}_1, \ldots, \mathrm{X}_n$ *des variables aléatoires. Alors*

$$\mathrm{E}(\mathrm{X}_1 + \ldots + \ldots + \mathrm{X}_n) = \mathrm{E}(\mathrm{X}_1) + \ldots + \mathrm{E}(\mathrm{X}_n)$$

Si ces variables sont indépendantes, on a de plus :

$$\mathrm{Var}(\mathrm{X}_1 + \ldots + \ldots + \mathrm{X}_n) = \mathrm{Var}(\mathrm{X}_1) + \ldots + \mathrm{Var}(\mathrm{X}_n)$$

Démonstration Soient $\mathrm{E}_1, \ldots, \mathrm{E}_n$ les ensembles de valeurs possibles des variables $\mathrm{X}_1, \ldots, \mathrm{X}_n$. Pour $n = 2$, on a

$$\begin{aligned}
\mathrm{E}(\mathrm{X}_1 + \mathrm{X}_2) &= \sum_{(x_1,x_2)\in \mathrm{E}_1\times \mathrm{E}_2} (x_1 + x_2)\,\mathrm{P}(\mathrm{X}_1 = x_1, \mathrm{X}_2 = x_2) \\
&= \sum_{(x_1,x_2)\in \mathrm{E}_1\times \mathrm{E}_2} x_1 \mathrm{P}(\mathrm{X}_1 = x_1, \mathrm{X}_2 = x_2) \\
&\quad + \sum_{(x_1,x_2)\in \mathrm{E}_1\times \mathrm{E}_2} x_2 \mathrm{P}(\mathrm{X}_1 = x_1, \mathrm{X}_2 = x_2) \\
&= \sum_{x_1\in \mathrm{E}_1} x_1 \mathrm{P}(\mathrm{X}_1 = x_1) \sum_{x_2\in \mathrm{E}_2} \mathrm{P}(\mathrm{X}_2 = x_2/\mathrm{X}_1 = x_1) \\
&\quad + \sum_{x_2\in \mathrm{E}_2} x_2 \mathrm{P}(\mathrm{X}_2 = x_2) \sum_{x_1\in \mathrm{E}_1} \mathrm{P}(\mathrm{X}_1 = x_1/\mathrm{X}_2 = x_2) \\
&= \mathrm{E}(\mathrm{X}_1) \times 1 + 1 \times \mathrm{E}(\mathrm{X}_2)
\end{aligned}$$

Le cas de n variables s'obtient alors par une récurrence immédiate.

Pour montrer la propriété d'additivité des variances, on commence par montrer le résultat suivant : Si X_1 et X_2 sont deux variables indépendantes, on a $\mathrm{E}(\mathrm{X}_1\mathrm{X}_2) = \mathrm{E}(\mathrm{X}_1)\,\mathrm{E}(\mathrm{X}_2)$. En effet,

$$\begin{aligned}
\mathrm{E}(\mathrm{X}_1\mathrm{X}_2) &= \sum_{(x_1,x_2)\in \mathrm{E}_1\times \mathrm{E}_2} x_1 x_2 \mathrm{P}(\mathrm{X}_1 = x_1, \mathrm{X}_2 = x_2) \\
&= \sum_{x_1\in \mathrm{E}_1} \sum_{x_2\in \mathrm{E}_2} x_1 x_2 \mathrm{P}(\mathrm{X}_1 = x_1)\,\mathrm{P}(\mathrm{X}_2 = x_2) \\
&= \sum_{x_1\in \mathrm{E}_1} x_1 \mathrm{P}(\mathrm{X}_1 = x_1) \sum_{x_2\in \mathrm{E}_2} x_2 \mathrm{P}(\mathrm{X}_2 = x_2) \\
&= \mathrm{E}(\mathrm{X}_1)\,\mathrm{E}(\mathrm{X}_2)
\end{aligned}$$

On pose alors $\mathrm{Y}_1 = \mathrm{X}_1 - \mathrm{E}(\mathrm{X}_1)$ et $\mathrm{Y}_2 = \mathrm{X}_2 - \mathrm{E}(\mathrm{X}_2)$. Les variables Y_1 et Y_2 sont encore indépendante, et l'on a $\mathrm{E}(\mathrm{Y}_1) = \mathrm{E}(\mathrm{Y}_2) = 0$. De ce qui précède, on déduit que $\mathrm{E}(\mathrm{Y}_1\mathrm{Y}_2) = 0$. D'où :

$$\begin{aligned}
\mathrm{Var}(\mathrm{X}_1 + \mathrm{X}_2) &= \mathrm{E}\left((\mathrm{Y}_1 + \mathrm{Y}_2)^2\right) \\
&= \mathrm{E}\left(\mathrm{Y}_1^2 + 2\mathrm{Y}_1\mathrm{Y}_2 + \mathrm{Y}_2^2\right) \\
&= \mathrm{E}\left(\mathrm{Y}_1^2\right) + 2\mathrm{E}(\mathrm{Y}_1\mathrm{Y}_2) + \mathrm{E}\left(\mathrm{Y}_2^2\right) \quad \text{(linéarité de l'espérance)} \\
&= \mathrm{Var}(\mathrm{X}_1) + \mathrm{Var}(\mathrm{X}_2) \quad (\text{car } \mathrm{E}(\mathrm{Y}_1\mathrm{Y}_2) = 0)
\end{aligned}$$

Le cas de n variables se traite de la même façon, mais les notations sont un peu plus lourdes..

□

1.3 Loi des grands nombres

Soient $\mathrm{X}_1, \ldots, \mathrm{X}_n$ des variables aléatoires indépendantes de même loi (c'est-à-dire, pour des variables discrètes, qui prennent les mêmes valeurs avec les mêmes probabilités, respectivement). Soit m leur espérance commune, σ^2 leur variance.

On pose
$$\overline{\mathrm{X}} = \frac{\mathrm{X}_1 + \ldots + \mathrm{X}_n}{n}$$

D'après ce qui précède, on a :

$$\begin{aligned}
\mathrm{E}\left(\overline{\mathrm{X}}\right) &= \frac{1}{n}\left(\mathrm{E}(\mathrm{X}_1) + \ldots \mathrm{E}(\mathrm{X}_n)\right) \\
&= \frac{1}{n}(m + \ldots + m) \\
&= m
\end{aligned}$$

et
$$\begin{aligned}
\mathrm{Var}\left(\overline{\mathrm{X}}\right) &= \frac{1}{n^2}\left(\mathrm{Var}(\mathrm{X}_1) + \ldots + \mathrm{Var}(\mathrm{X}_n)\right) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

En écrivant l'inégalité de Bienaymé-Chebychev pour la variable aléatoire $\overline{\mathrm{X}}$, on obtient le :

Théorème 1.4 *(Loi faible des grands nombres)*
Pour tout $a > 0$, on a :

$$\mathrm{P}\left(\left|\overline{\mathrm{X}} - m\right| \geq a\right) \leq \frac{\sigma^2}{na^2}$$

En particulier, lorsque n tend vers l'infini, c'est-à-dire lorsque la taille de l'échantillon considéré tend elle-même vers l'infini, la probabilité pour que la moyenne observée s'écarte de l'espérance tend vers 0.

1.4 Fiabilité du sondage

Dans le cas de notre sondage, nous avons déjà fait implicitement un certain nombre d'hypothèses : l'échantillon sondé est pris “au hasard”, c'est-à-dire que l'on suppose les n réponses indépendantes (les personnes interrogées ne s'influencent pas). Et l'on considère que chaque individu a une probabilité p de répondre “OUI”, $1-p$ de répondre “NON”, où le nombre p ne dépend pas de l'individu.

Dans notre modèle, les réponses des n personnes interrogées peuvent être modélisée par n variables aléatoires indépendantes $\mathrm{X}_1, \ldots, \mathrm{X}_n$, à valeur dans l'ensemble $\{0,1\}$, ayant toutes pour loi la loi de Bernoulli de paramètre p.

L'espérance de cette loi vaut $\mathrm{E}(\mathrm{X}) = p \times 1 + (1-p) \times 0 = p$, et sa variance vaut $\sigma^2 = \mathrm{E}\left((\mathrm{X}-p)^2\right) = p \times (1-p)^2 + (1-p) \times p^2 = p\,(1-p)$.

En appliquant la loi des grands nombres, on a donc, si l'on note $\overline{\mathrm{X}}$ la variable aléatoire $\dfrac{\mathrm{X}_1 + \ldots + \mathrm{X}_n}{n}$, la proportion de personnes répondant “OUI” est la réalisation de la variable $\overline{\mathrm{X}}$, et pour a un réel positif :

$$\mathrm{P}\left(\left|\overline{\mathrm{X}} - p\right| \geq a\right) \leq \frac{p\,(1-p)}{na^2} \leq \frac{1}{4na^2}$$

(On a utilisé la majoration $p\,(1-p) \leq \dfrac{1}{4}$ pour tout réel p de l'intervalle $[\,0\,;1\,]$.)

Si l'on s'intéresse à la question de savoir si le référendum sera ou non accepté, par exemple, la proportion de réponse positives étant de $0,6$, pour 1000 personnes interrogées, la marge d'erreur que l'on se donne sera ici $a = 0,1$.

$$\mathrm{P}\left(\left|\overline{\mathrm{X}} - p\right| \geq 0,1\right) \leq \frac{1}{4000\,(0,1)^2} = \frac{1}{40}$$

Autrement dit, si notre modèle est le bon, ce qui est fort légitime, il y a moins de $2,5\%$ de chances pour que la proportion de sondés votant “OUI” soit en dehors de l'intervalle $[\,p-0,1\,;p+0,1\,]$.

2 Intervalles de confiance

Nous arrivons à une idée essentielle de l'inférence statistique, celle d' “intervalle de confiance”. C'est un intervalle aléatoire, qui dépend de la réalisation observée. Nous cherchons à construire un intervalle qui contienne la vraie valeur de p avec une probabilité élevée, tout en gardant l'intervalle aussi petit que possible. En effet, on peut par exemple dire avec une probabilité 1 que $p \in [\,0\,;1\,]$, mais l'information contenue dans cette affirmation est, pour le moins, faible.

Soit C une fonction définie sur l'intervalle $[\,0\,;1\,]$, qui à tout réel associe un sous-intervalle de $[\,0\,;1\,]$.

Définition Soit $0 < \alpha < 1$. L'application C est appelée intervalle de confiance pour p au risque α (ou au niveau $1-\alpha$), si pour tout $p \in [\,0\,;1\,]$ on a

$$\mathrm{P}_p\left(p \in \mathrm{C}\left(\overline{\mathrm{X}}\right)\right) \geq 1-\alpha$$

N.B. La notation $\mathrm{P}_p(\mathrm{A})$ désigne la probabilité de l'événement A lorsque X a une probabilité p de dire “OUI”. Pour d'autres types de variables aléatoires, on peut avoir à fixer plus de paramètres pour donner un sens à cette expression, mais ici p suffit à caractériser la "loi" des variables de Bernoulli considérées.

Ainsi, si $\overline{\mathrm{X}}$ désigne la proportion observée, ici la proportion des sondés répondant "OUI", l'on affirme que la proportion réelle p de la population favorable au référendum est dans l'intervalle $\mathrm{C}\left(\overline{\mathrm{X}}\right)$. Cette affirmation est bien sûr sujette à caution : on peut se tromper, mais la probabilité d'une telle éventualité est faible (inférieure à α).

Autrement dit, la proportion réelle, inconnue, étant précisément l'objet de l'inférence, il nous faut considérer toutes les valeurs possibles de p : le risque α est un réel tel que, à priori, quelle que soit la valeur p, on a une probabilité faible (plus petite que α) pour que p ne soit pas dans l'intervalle aléatoire $\mathrm{C}\left(\overline{\mathrm{X}}\right)$ construit à partir de la proportion observée $\overline{\mathrm{X}}$.

Dans le cadre de notre problème, voyons comment construire un intervalle de confiance pour p, au niveau α. On a supposé que les réponses des personnes interrogées étaient indépendantes entre elles, et donc le nombre de réponses "OUI", que nous désignerons par la variable aléatoire $\mathrm{X}_. = \mathrm{X}_1 + \ldots + \mathrm{X}_n$, suit une loi binomiale :

$$k \mapsto b\left(k, n, p\right) = \mathrm{C}_n^k p^k \left(1-p\right)^{n-k}$$

N.B. Les variables $\mathrm{X}_.$ et $\overline{\mathrm{X}}$ sont fortement corellée : nous avons $\overline{\mathrm{X}} = \dfrac{\mathrm{X}_.}{n}$. La connaissance de l'une étant équivalente à la connaissance de l'autre, nous pouvons baser notre construction sur l'une où l'autre indifféremment.

Lemme 2.1 *Soit k un entier. La fonction* B

$$p \mapsto \sum_{i=0}^{k} b\left(i, n, p\right) = \mathrm{P}_p\left(\mathrm{X}_. \leq k\right)$$

est décroissante, strictement si $k < n$.

Ce résultat est assez intuitif : si pour chaque personne interrogée, la probabilité de réponse "OUI" augmente, alors la probabilité pour que le nombre total de réponse "OUI" soit plus petit qu'une borne fixée diminue !

Démonstration Nous allons tout simplement calculer la dérivée de B.

$$\begin{aligned}
\mathrm{B}'\left(p\right) &= \sum_{i=1}^{k} \mathrm{C}_n^i i p^{i-1} \left(1-p\right)^{n-i} - \sum_{i=0}^{k} \mathrm{C}_n^i \left(n-i\right) p^i \left(1-p\right)^{n-i-1} \\
&= n \sum_{i=1}^{k} \mathrm{C}_{n-1}^{i-1} p^{i-1} \left(1-p\right)^{n-i} - n \sum_{i=0}^{k} \mathrm{C}_{n-1}^{i} p^i \left(1-p\right)^{n-1-i} \\
&= -n\mathrm{C}_{n-1} k p^k \left(1-p\right)^{n-1-k}
\end{aligned}$$

Cette dérivée est bien négative, donc la fonction est décroissante. On en déduit que la fonction $p \mapsto \mathrm{P}_p\left(\mathrm{X}_. \geq k\right)$ est elle croissante, ce qui est également naturel.

□

Soit k un entier compris entre 0 et n.

Comme la fonction $p \mapsto \mathrm{P}_p\left(\mathrm{X}_. \geq k\right)$ est croissante, on peut définir un réel

$$p_1\left(k\right) = \inf\{p, \mathrm{P}_p\left(\mathrm{X}_. \geq k\right) > \frac{\alpha}{2}\}$$

De la même façon, on définit un réel

$$p_2(k) = \sup\{p, \mathrm{P}_p(\mathrm{X}_. \leq k) > \frac{\alpha}{2}\}$$

Théorème 2.2 *Ayant observé la valeur de k pour la variable $\mathrm{X}_.$, l'intervalle :*

$$\mathrm{C}(k) = \left]\, p_1(k)\ ; p_2(k) \right[$$

est alors un intervalle de confiance pour p de niveau $1-\alpha$,

Démonstration Il nous faut montrer que, quel que soit p, on a :

$$\mathrm{P}_p(p \in \mathrm{C}(\mathrm{X}_.)) \geq 1-\alpha$$

On a en effet d'après la définition de $p_2(k)$ et la décroissance de la fonction $p \mapsto \mathrm{P}_p(\mathrm{X}_. \leq k)$:

$$\mathrm{P}_p(\mathrm{X}_. \leq k) > \frac{\alpha}{2} \Longleftrightarrow p < p_2(k)$$

de même $$\mathrm{P}_p(\mathrm{X}_. \geq k) > \frac{\alpha}{2} \Longleftrightarrow p > p_1(k)$$

Or à p fixé, la suite $k \mapsto \mathrm{P}_p(\mathrm{X}_. \leq k)$ est bien entendu croissante, elle vaut 0 pour $k=-1$ et 1 pour $k=n$. Il existe donc un entier $n_1(p)$ tel que :

$$\mathrm{P}_p(\mathrm{X}_. \leq k) > \frac{\alpha}{2} \Longleftrightarrow k \geq n_1(p)$$

De la même façon, il existe un entier $n_2(p)$ tel que :

$$\mathrm{P}_p(\mathrm{X}_. \geq k) > \frac{\alpha}{2} \Longleftrightarrow k \leq n_2(p)$$

On en déduit les équivalences :

$$p < p_2(k) \Longleftrightarrow k \geq n_1(p)$$

et $$p > p_1(k) \Longleftrightarrow k \leq n_2(p)$$

c'est-à-dire $$p \in \mathrm{C}(k) \Longleftrightarrow n_1(p) \leq k \leq n_2(p)$$

Mais les deux entiers $n_1(p)$ et $n_2(p)$ sont tels que :

$$\begin{aligned}\mathrm{P}_p(n_1(p) \leq \mathrm{X}_. \leq n_2(p)) &= 1 - \mathrm{P}_p(\mathrm{X}_. \leq n_1(p)-1) - \mathrm{P}_p(\mathrm{X}_. \geq n_2(p)+1) \\ &\geq 1 - \frac{\alpha}{2} - \frac{\alpha}{2}\end{aligned}$$

On obtient :

$$\mathrm{P}_p(p \in \mathrm{C}(\mathrm{X}_.)) = \mathrm{P}_p(n_1(p) \leq \mathrm{X}_. \leq n_2(p)) \geq 1-\alpha$$

C'est-à-dire que l'intervalle $\mathrm{C}(\mathrm{X}_.)$ est bien un intervalle de confiance pour p de niveau $1-\alpha$.

□

N.B. Cet intervalle est un des plus précis qui soient, puisque l'on ne fait pour le construire aucune approximation. En revanche, sa construction effective demanderait des calculs assez lourds, et nécessitant de grosses puissances de calculs pour des échantillons suffisament grands. En pratique, on utilise alors pour construire un intervalle de confiance des approximation par des lois connues et plus faciles à manipuler (loi de Poisson ou loi Gaussienne). Pour de petits échantillons, on se réfère à des tables de valeurs de probabilités binomiales cumulées.

Revenons au cas de notre sondange, et voyons quel sera notre intervalle de confiance en fonction de la taille de l'échantillon.

Par exemple, si l'on a interrogé 10 personnes, et que 6 ont répondu "OUI", on obtient comme intervalle de confiance de niveau 95% l'intervalle $[0,262 ; 0,878]$. On dit alors que la proportion réelle p est dans cet intervalle, ce avec une probabilité d'erreur de 5%. Mais cet intervalle est trop large pour nous, ce qui est normal puisque l'on a pris un échantillon très restreint.

Si l'on interroge 100 personnes, dont 60 répondent "OUI", les tables de valeurs numériques nous donneront pour intervalle de confiance de niveau 95% l'intervalle $[0,497 ; 0,697]$. On peut ici affirmer qu'il y a donc presque 95% de chances pour que le référendum soit accepté.

Pour des échantillons plus conséquents, les tables de valeurs ne suffisent plus, la plupart du temps. Celles-ci ne sont en effet pas infinies, et de toutes façons, lorsque n est suffisamment grand, on peut utiliser des approximations de nos variables par des gaussiennes. En l'occurence pour 1000 personnes sondées, la variable aléatoire $\overline{\mathrm{X}} = \dfrac{\mathrm{X}_1 + \ldots + \mathrm{X}_{1000}}{1000}$ suit une loi très proche de la loi gaussienne de moyenne p et de variance $\dfrac{p\,(1-p)}{1000}$. Si le nombre de réponses positives est de 600, un intervalle de confiance pour p de niveau 95% sera donc obtenu en cherchant la valeur $p_1\,(0,6)$, valeur minimale de p pour qu'une gaussienne de moyenne p et de variance $\dfrac{p\,(1-p)}{1000}$ soit supérieure ou égale à $0,6$ avec une probabilité d'au moins $2,5\%$; et la valeur $p_2\,(0,6)$, valeur maximale de p pour qu'une gaussienne de moyenne p et de variance $\dfrac{p\,(1-p)}{1000}$ soit inférieur ou égale à $0,6$ avec une probabilité d'au moins $2,5\%$. Pour trouver ces valeurs, on se ramène à des conditions sur une gaussienne de moyenne 0 et de variance 1, ce qui nous donne d'après la table de valeurs numériques :

$$p_1\,(0,6) \text{ et } p_2\,(0,6) \text{ racines de } \quad \left| \frac{0,6-p}{\sqrt{\dfrac{p\,(1-p)}{1000}}} \right| = 2$$

On trouve alors $p_1\,(0,6) \simeq 0,5686$ et $p_2\,(0,6) \simeq 0,6305$. Notre intervalle de confiance pour p, de niveau 95%, sera donc l'intervalle $[0,569 ; 0,631]$.

N.B. En language courant, on dira que ce sondage a une marge d'erreur d'un peu plus de 3%. C'est du moins ce que l'on entend de temps en temps à propos de sondages effectués, comme par hasard, auprès de 1015 personnes ! Pour un tel échantillon, la longueur maximale de l'intervalle de confiance est de l'ordre de $6,14\%$. (Pour avoir une marge d'erreur maximale de moins de 3%, il faut un échantillon d'au moins 1064 personnes.)

A Approximation de lois binomiales

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de même loi $\mathcal{L}$. Si $x_1, \ldots, x_n$ sont des réalisations des variables $X_1, \ldots, X_n$, on dit que c'est un n-échantillon de la loi $\mathcal{L}$.

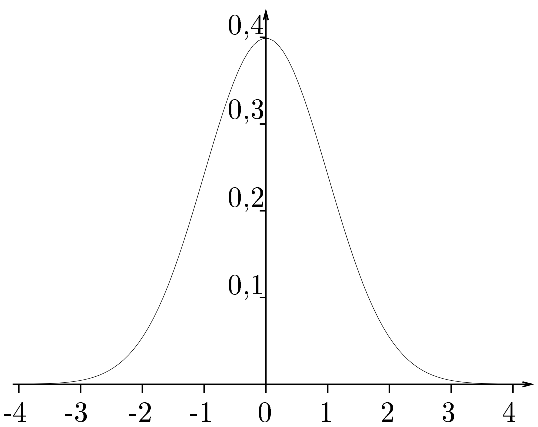
Supposons maintenant que $X_1, \ldots, X_n$ est un n-échantillon de la loi de Bernoulli de paramètre p, c'est-à-dire que chacun des X_i vaut 1 avec la probabilité p et 0 avec la probabilité $1-p$. Alors leur somme $S = X_1 + \ldots + X_n$ suit une loi binomiale $\mathcal{B}(n,p)$.

Définition Une variable aléatoire réelle Y est dite suivre la loi Gaussienne centrée réduite $\mathcal{N}(0,1)$, d'espérance nulle et de variance 1, si pour tous a et b réels ($a < b$), la probabilité pour que Y soit dans l'intervalle $[\,a\,;b\,]$ est égale à :

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\, \mathrm{d}x$$

N.B. Une telle variable aléatoire est dite "à densité" : elle peut prendre pour valeur tout réel. Elle a pour densité la fonction $x \mapsto e^{-\frac{x^2}{2}}/\sqrt{2\pi}$, c'est-à-dire que la probabilité $\mathrm{P}(\mathrm{Y} \in [\,x\,;x + \mathrm{d}x\,])$ est, lorsque la longueur $\mathrm{d}x$ est très petite, proportionnelle à $\mathrm{d}x$, et est donnée par $e^{-\frac{x^2}{2}}/\sqrt{2\pi}\,\mathrm{d}x$.

N.B. C'est la fameuse courbe "en cloche" : la fonction $x \mapsto e^{-\frac{x^2}{2}}/\sqrt{2\pi}$ a pour courbe représentative :

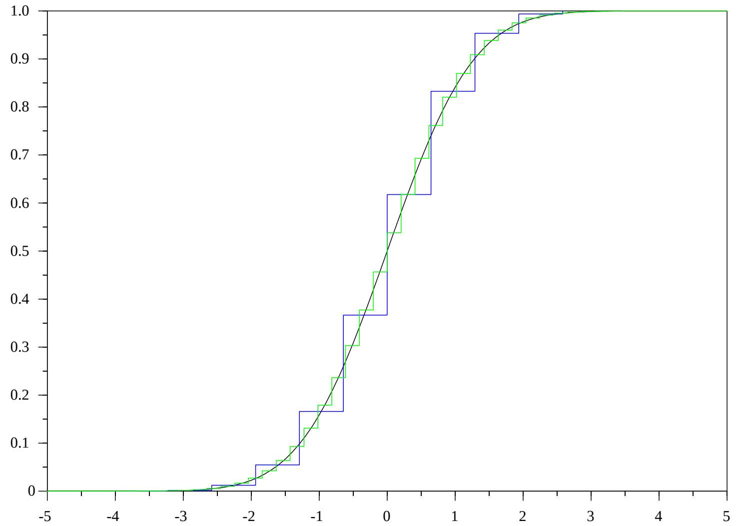


Propriété A.1 *Lorsque n est grand, et que np et $n(1-p)$ ne sont pas trop petit, alors la loi de*

$$\frac{S - np}{\sqrt{np\,(1-p)}}$$

est proche d'une loi Gaussienne centrée réduite $\mathcal{N}(0,1)$.

La figure suivante illustre ce phénomène : elle représente les fonctions de répartition de $\dfrac{S - np}{\sqrt{np\,(1-p)}}$, toujours pour $p = 0,6$, et pour des valeurs de n de 10 (courbe bleue) et 100 (courbe verte). Celles-ci se rapprochent peu à peu de la courbe noire, qui représente la fonction de répartition d'une variable aléatoire gaussienne ($x \mapsto \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, \mathrm{d}t$).



N.B. La fonction de répartition d'une variable aléatoire X est la fonction de $\mathbb{R}$ dans $[\,0\,;1\,]$, qui à un réel x associe la probabilité $\mathrm{P}\,(\mathrm{X} \leq \mathrm{X})$.

La démonstration de cette convergence est assez technique. On peut en trouver une version assez accessible dans [2].

En fait, ce phénomène est un cas particulier d'un phénomène plus général :

Théorème A.2 *(Théorème central limite)*

Soit $\mathrm{X}_1, \dots, \mathrm{X}_n$ *un* n*-échantillon d'une loi* $\mathcal{L}$ *de moyenne* m *et de variance* σ^2*. On pose* $\mathrm{S} = \mathrm{X}_1 + \ldots + \mathrm{X}_n$*. Alors, lorsque* n *tend vers l'infini, la loi de la variable :*

$$\frac{\mathrm{S} - nm}{\sigma\sqrt{n}}$$

converge vers la loi Gaussienne centrée réduite $\mathcal{N}\,(0,1)$*.*

En d'autres termes, lorsque n est suffisamment grand, tout se passe comme si la loi $\mathcal{L}$ des X_i était la loi Gaussienne $\mathcal{N}\left(m, \sigma^2\right)$. La loi de S/n est alors très proche de $\mathcal{N}\left(m, \sigma^2/n\right)$.

La loi gaussienne n'est pas la seule loi possible pour approximer une loi binomiale. Lorsque n est grand et $\lambda = np$ est petit, la loi de S est proche d'une loi de Poisson de paramètre λ, c'est-à-dire que pour k un entier, on a $\mathrm{P}_p\,(\mathrm{S} = k) \sim e^{-\lambda}\dfrac{\lambda^k}{k!}$.

Références

[1] K. Krickeberg, *Petit cours de statistique*, Springer, 1996.

[2] D. Dacunha-Castelle & M. Duflo, *Probabilités et statistiques, tome 1*, Masson, 1982 Mathématiques appliquées pour la maîtrise.

[3] M. Lavieville, *Statistiques et probabilités : rappels de cours et exercices corrigés*, Dunod, 1996.