Fluctuation, prise de décision, estimation

Charles SUQUET

http://math.univ-lille1.fr/~suquet/

Université Lille 1 – Sciences et Technologies
CNRS UMR 8524

Table des matières

1 Concentration de la mesure

Le phénomène de concentration de la mesure (ou de la loi de probabilité d'une variable aléatoire) permet en statistique, d'obtenir de l'information, voire des quasi certitudes, à partir d'observations aléatoires. Pour l'illustrer graphiquement, comparons les diagrammes en bâtons de la loi uniforme sur $[\![0, 100]\!]$ et de la loi binomiale de paramètres 100 et 0,3 représentés figure 1. Rappelons que dans le diagramme en bâtons de la loi d'une variable aléatoire discrète X, le segment vertical d'abscisse x_k (ou « bâton ») a pour hauteur $P(X = x_k)$. Pour les deux lois considérées ici, il y a théoriquement un bâton de hauteur non nulle pour chacune des valeurs entières $x_k = k \in [\![0, 100]\!]$. Pour la loi uniforme, tous les bâtons ont la même hauteur $1/101$, tandis que pour la loi binomiale, les bâtons d'abscisse en dehors de l'intervalle $[\![12, 48]\!]$ ont une hauteur trop petite pour être visibles. On peut d'ailleurs vérifier par le calcul que la somme des hauteurs de tous ces bâtons « invisibles » vaut

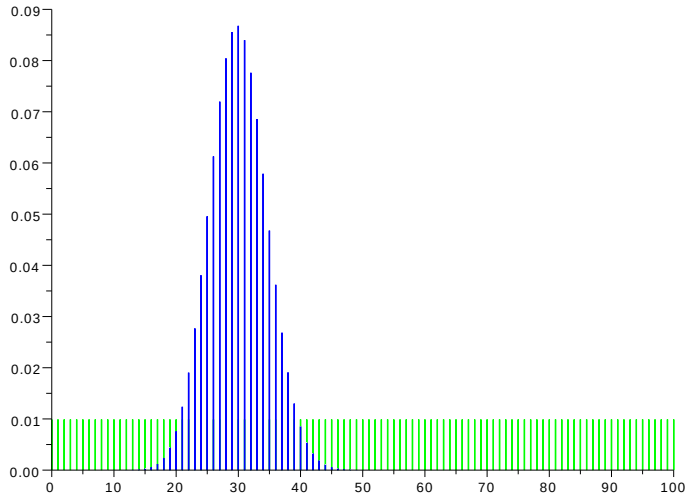

FIGURE 1 – Diagrammes en bâtons des lois $\mathrm{Bin}(100; 0,3)$ et $\mathrm{Unif}(\llbracket 0, 100 \rrbracket)$

0,000 057 5. Autrement dit, si la variable aléatoire S suit la loi binomiale de paramètres 100 et 0,3, $P(12 \leqslant S \leqslant 48) = 0{,}999\,942\,5$. Par contre, si U suit la loi uniforme sur $\llbracket 0, 100 \rrbracket$, $P(12 \leqslant U \leqslant 48) = 37/101 \simeq 0{,}366\,4$. Si avant d'observer une valeur de S ou de U, on parie qu'elle va tomber dans $\llbracket 12, 48 \rrbracket$, on est pratiquement sûr de gagner avec S et on a presque 2 chances sur 3 de perdre avec U. C'est cette concentration de la probabilité sur un intervalle court (relativement à la longueur du support) pour S qui permet de faire de l'estimation ou de la prise de décision à partir des observations.

Ce phénomène de concentration de la loi binomiale s'accentue quand n augmente. Pour le visualiser, regardons le diagramme en bâtons d'une loi binomiale avec $p = 0{,}3$ pour de grandes valeurs de n. On se contentera ici de $n = 1000$ (on peut encore calculer simplement les coefficients binomiaux pour cette valeur en Scilab, en utilisant le triangle de Pascal).

- Sur $\llbracket 0, n \rrbracket$, cf figure 2, ceci illustre la *loi faible des grands nombres*, à condition de faire par la pensée une mise à l'échelle en $1/n$, ce qui revient ici à décréter que la graduation horizontale 1000 correspond à 1.
- Sur $[np - 4\sqrt{np(1-p)}, np + 4\sqrt{np(1-p)}]$, cf figure 3, ceci illustre le théorème de de Moivre Laplace (cas particulier du *théorème limite central*).

Notons que le passage de la figure 2 à la figure 3 n'est rien d'autre qu'un changement d'échelle horizontale. Le comportement mathématique que j'es-

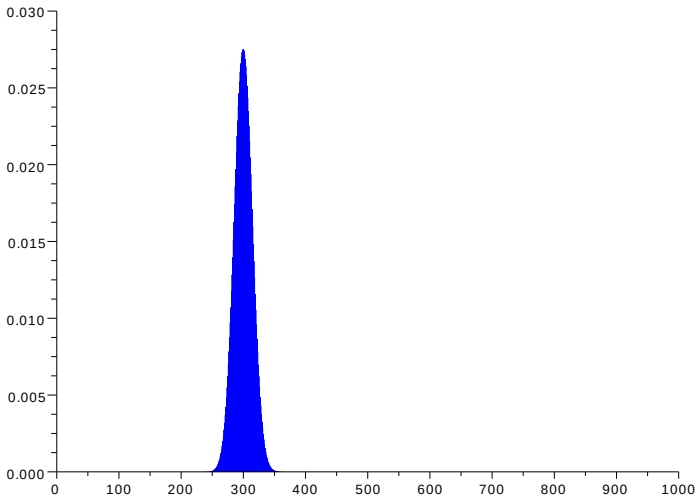

FIGURE 2 – Diagrammes en bâtons de $\mathrm{Bin}(1000; 0,3)$, loi faible des grands nombres

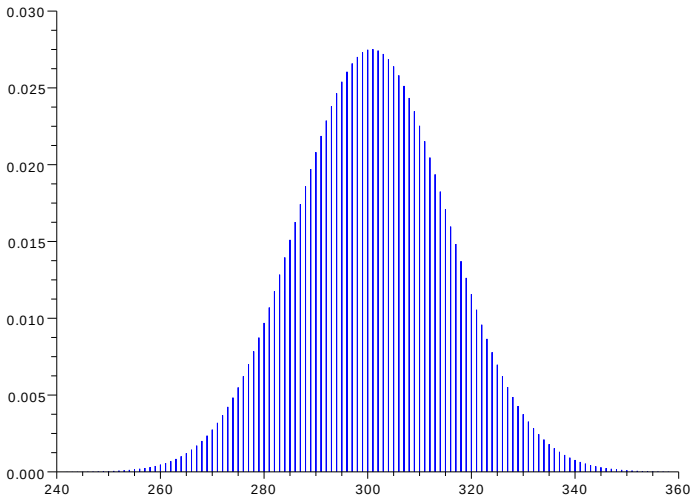

FIGURE 3 – Diagrammes en bâtons de $\mathrm{Bin}(1000; 0,3)$, théorème limite central

père illustrer par ces figures est le suivant. Si S_n est une variable aléatoire de loi binomiale de paramètres n et p, alors S_n/n converge *en probabilité* vers p quand n tend vers l'infini. C'est l'exemple le plus simple de loi faible des grands nombres et qui est la justification de l'estimation d'une probabilité inconnue par la fréquence de réalisation observée sur un grand échantillon. De manière équivalente, on peut aussi dire que $S_n/n - p$ converge en probabilité vers 0. En pratique, il est utile d'avoir une idée de la vitesse de convergence. Pour cela, on regarde le comportement de $Z_n = c_n(S_n/n - p)$ où (c_n) est une suite de constantes (i.e. non aléatoires) tendant vers l'infini. Intuitivement, si c_n tend trop lentement vers $+\infty$, cela ne va rien changer et Z_n convergera vers 0 en probabilité. Si c_n tend trop vite vers l'infini, il y aura explosion avec oscillation indéfinie de Z_n entre $-\infty$ et $+\infty$ par amplification des fluctuations aléatoires de S_n/n autour de p. Il se trouve qu'il y a une situation intermédiaire quand c_n est de l'ordre de grandeur de $\sqrt{n}$, où Z_n ne converge plus en probabilité vers 0, mais n'explose pas pour autant. La loi de Z_n reste pour l'essentiel concentrée sur un intervalle de taille constante. Il y a en fait *convergence en loi* de Z_n vers une variable [1] Z de loi gaussienne de paramètres 0 et $\sigma = \sqrt{p(1-p)}$. Pour une telle Z, $P(-4\sigma \leqslant Z \leqslant 4\sigma) \simeq 0{,}999\,946\,7$. En revenant au comportement de S_n plutôt qu'à celui de S_n/n (cela revient à tout multiplier par n) ceci explique pourquoi le choix de « zoomer » sur l'intervalle $\left[np - 4\sqrt{np(1-p)}, np + 4\sqrt{np(1-p)}\right]$ a permis de capturer et visualiser l'essentiel de la masse de la loi de S_n.

Voici l'énoncé précis sur la convergence de Z_n (en fait de Z_n/σ).

Théorème 1 (de de Moivre-Laplace)**.** *Si S_n est une variable aléatoire de loi binomiale de paramètres n et $p \in]0,1[$, on a avec $q := 1-p$,*

$$S_n^* := \frac{S_n - np}{\sqrt{npq}} = \sqrt{\frac{n}{pq}}\left(\frac{S_n}{n} - p\right) \xrightarrow[n\to+\infty]{\text{loi}} Z,$$

où Z est une variable de loi gaussienne $\mathfrak{N}(0,1)$.

« Traduction [2] » : pour tous réels a, b avec $a < b$,

$$P\big(S_n^* \in I(a,b)\big) \xrightarrow[n\to+\infty]{} \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(\frac{-t^2}{2}\right) \mathrm{d}t,$$

1. L'expression convergence en loi d'une suite de *variables* aléatoires est un grossier (mais usuel) abus de langage. En fait, c'est de convergence de la *loi* de Z_n qu'il s'agit. Z_n converge tout aussi bien en loi vers n'importe quelle autre variable aléatoire Z' ayant même loi que Z.

2. La convergence en loi d'une suite de variables aléatoires (Z_n) vers une variable aléatoire Z peut se définir comme la convergence de la suite (F_n) des fonctions de répartition vers la fonction de répartition F de Z, *en tout point de continuité de F*. Lorsque F est continue sur $\mathbb{R}$, ce qui est le cas pour la loi $\mathfrak{N}(0,1)$, cette définition se simplifie sensiblement.

où $I(a, b)$ est n'importe lequel des quatre intervalles d'extrémités a et b.

Le théorème de de Moivre-Laplace est historiquement le premier exemple de théorème de convergence en loi vers une gaussienne. Aujourd'hui, il est considéré comme un corollaire du théorème suivant.

Théorème 2 (théorème limite central)**.** *Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$, indépendantes, de même loi et de carré intégrable (et non* p.s. *constantes*[3]*). Notons $\mu := \mathbf{E}\, X_1$, $\sigma^2 := \operatorname{Var} X_1$ avec $\sigma > 0$ et $S_n = \sum_{k=1}^{n} X_k$. Définissons la somme centrée réduite :*

$$S_n^* := \frac{S_n - \mathbf{E}\, S_n}{\sqrt{\operatorname{Var} S_n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mu\right).$$

Alors

$$S_n^* \xrightarrow[n\to+\infty]{\text{loi}} Z,$$

où Z est une variable de loi gaussienne $\mathfrak{N}(0, 1)$.

Pour voir que le théorème de de Moivre-Laplace est une application immédiate du théorème limite central, il suffit de prendre les X_i de loi de Bernoulli de paramètre p. S_n suit alors la loi binomiale de paramètres n et p et a pour espérance np et pour variance npq.

2 Prise de décision

Commençons par une citation du programme de mathématiques pour la classe de seconde (2009–2010).

> « L'intervalle de fluctuation au seuil de 95%, relatif aux échantillons de taille n, est l'intervalle centré autour de p, proportion du caractère dans la population, où se situe, avec une probabilité égale à 0,95, la fréquence observée dans un échantillon de taille n. Cet intervalle peut être obtenu, de façon approchée, par simulation. Le professeur peut indiquer aux élèves le résultat suivant, utilisable dans la pratique pour des échantillons de taille $n \geqslant 25$ et des proportions p du caractère comprises entre 0,2 et 0,8 : si f désigne la fréquence du caractère dans l'échantillon, f appartient à l'intervalle $[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}]$ avec une probabilité d'au moins 0,95. Le professeur peut faire percevoir expérimentalement la validité de cette propriété mais elle n'est pas exigible. »

3. Si X_i est presque-sûrement constante, alors $P(X_i = c) = 1$ pour une certaine constante c et dans ce cas $\operatorname{Var} X_i = 0$. La réciproque est vraie.

L'intérêt pédagogique de cette notion d'intervalle de fluctuation est de fournir un outil simple pour une première approche de la prise de décision statistique. Même si la définition ci-dessus n'est pas totalement satisfaisante pour un mathématicien [4], elle donne d'assez bons résultats en pratique et permet d'aller directement à l'essentiel : l'exploitation du phénomène de concentration de la mesure (ici de la loi binomiale) comme aide statistique à la prise de décision. Illustrons le d'abord par un exemple un peu artificiel et volontairement simpliste.

Problème à deux urnes. On dispose de deux urnes d'apparence complètement identique, numérotées 1 et 2, le numéro étant masqué. On sait que l'urne 1 contient 30% de boules vertes et 70% de boules rouges tandis que l'urne 2 contient 78% de boules vertes et 22% de boules rouges. On en choisit une au hasard dans laquelle on effectue 100 tirages avec remise d'une boule. On note le nombre S d'apparitions des boules vertes lors de cette suite de tirages et on doit dire, au vu de ce nombre de boules vertes, si le numéro de l'urne choisie est ou non le 1.

Si les tirages ont lieu dans l'urne 1, la loi de S est la binomiale de paramètres 100 et 0,3, tandis qu'avec des tirages dans l'urne 2, la loi de S est la binomiale de paramètres 100 et 0,78. Il s'agit donc de mettre en concurrence ces deux lois possibles pour S, *au vu de l'observation d'une valeur de* S. Un coup d'oeil sur les diagrammes en bâtons (figure 4) montre qu'il n'y a pas photo et que même si ces deux lois ont le même support théorique à savoir $[\![0, 100]\!]$, en pratique elles ne vivent pas sur le même territoire. Les intervalles de fluctuation au seuil 95% sont (approximativement) $[\![20, 40]\!]$ pour la première et $[\![68, 88]\!]$ pour la seconde [5]. Au vu du graphique, on se convaincra facilement que pour l'urne 1, la probabilité que S prenne ses valeurs dans $[\![10, 50]\!]$ est « pratiquement » égale à 1, tandis que pour l'urne 2 il en va de même avec l'intervalle $[\![60, 95]\!]$. Quel que soit le numéro de l'urne choisie, il est donc très peu vraisemblable que la valeur de S observée soit en dehors de l'un de ces deux intervalles. Si elle est dans le premier, on pourra conclure avec un risque d'erreur infime que l'urne choisie était la numéro 1. Si S est dans $[\![60, 95]\!]$, on concluera de même que l'urne choisie était la numéro 2. En remplaçant ces deux intervalles par les intervalles de fluctuation au seuil 95%, la règle de décision sera la même, avec un risque d'erreur un peu plus grand, inférieur ou égal à 5%. Attention, j'ai écrit « risque » et non pas « probabilité », car il y a ici deux probabilités et deux types d'erreur possible, chacune pouvant être mesurée avec l'une de ces deux probabilités.

4. Voir la section 5.
5. Intervalles de fluctuation obtenus par la formule approchée $[np - \sqrt{n}, np + \sqrt{n}]$.

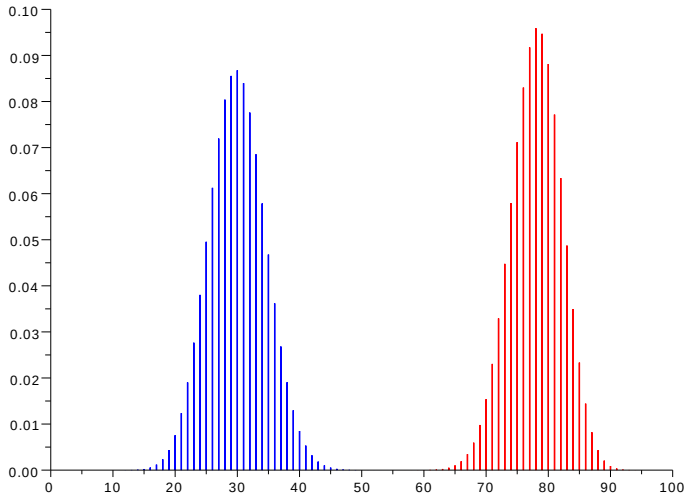

FIGURE 4 – Diagrammes en bâtons de $\mathrm{Bin}(100; 0,3)$ et $\mathrm{Bin}(100; 0,78)$

Test sur la valeur d'une probabilité. Une situation plus complexe, mais plus proche de problèmes réels, est celle où on a une seule urne et *des raisons de croire* que sa composition est (par exemple) de 30% de boules vertes et 70% de rouges. On effectue à nouveau 100 tirages et au vu de la valeur de S, on doit décider si on conserve ou rejette cette hypothèse sur la composition de l'urne. C'est un problème de test sur la valeur d'une proportion (ou plus généralement d'une probabilité). Là encore, on peut utiliser l'intervalle de fluctuation pour se donner une règle de décision. Puisque l'intervalle de fluctuation au seuil de 95% est $[\![20, 40]\!]$, l'observation d'une valeur de S « trop petite » (ici strictement inférieure à 20) ou « trop grande » (ici strictement supérieure à 40) conduira à rejetter l'hypothèse d'une composition d'urne à 30% de boules vertes. On peut parler ici de zone d'acceptation *bilatérale*.

Dans certaines situations, il est naturel de rejetter l'hypothèse faite sur une proportion (ou une probabilité), seulement lorsque S prend une valeur « trop grande » c'est le cas par exemple avec les problèmes de questionnaire à choix multiple. On prend comme hypothèse que le candidat a répondu au hasard à chaque question. Le nombre S de bonnes réponses est alors une variable aléatoire de loi binomiale avec pour paramètres n le nombre de questions, $p = c/d$ où d est le nombre de réponses proposées par question dont c sont correctes (le plus souvent, $c = 1$). On décidera alors de rejetter l'hypothèse si la valeur observée de S est supérieure ou égale à b, où b est

le plus petit entier tel que $P(S \leqslant b) > 0{,}95$. Ici la zone d'acceptation sera $[\![0, b-1]\!]$, c'est une zone unilatérale [6].

À propos de Q.C.M., voici un petit problème sur le Code de la Route, contenant une modélisation plus réaliste (attention, c'est une digression !).

Code de la Route I. Pour l'examen du Code de la Route, les candidats doivent remplir un questionnaire de 40 questions en choisissant pour chacune d'elles l'une des 4 réponses proposées, dont une seule est exacte. Tout candidat ayant obtenu au moins 36 bonnes réponses est déclaré reçu. Un candidat totalement ignorant décide de tenter sa chance en cochant complètement au hasard une réponse pour chaque question.

Le nombre S de bonnes réponses du candidat est ici le nombre de succès dans une suite de 40 épreuves répétées indépendantes, avec pour chacune probabilité de succès $1/4$. La variable aléatoire S suit donc la loi binomiale de paramètres 40 et $1/4$.

Pour calculer $P(S \geqslant 36)$, on utilise la décomposition

$$P(S \geqslant 36) = \sum_{k=36}^{40} P(S = k).$$

$$\begin{aligned} P(S \geqslant 36) &= C_{40}^{36}\Big(\frac{1}{4}\Big)^{36}\Big(\frac{3}{4}\Big)^{4} + C_{40}^{37}\Big(\frac{1}{4}\Big)^{37}\Big(\frac{3}{4}\Big)^{3} + C_{40}^{38}\Big(\frac{1}{4}\Big)^{38}\Big(\frac{3}{4}\Big)^{2} \\ &\quad + C_{40}^{39}\Big(\frac{1}{4}\Big)^{39}\Big(\frac{3}{4}\Big)^{1} + C_{40}^{40}\Big(\frac{1}{4}\Big)^{40}\Big(\frac{3}{4}\Big)^{0} \\ &= \frac{1}{4^{40}}\big(C_{40}^{36} \times 3^4 + C_{40}^{37} \times 3^3 + C_{40}^{38} \times 3^2 + C_{40}^{39} \times 3 + C_{40}^{40} \times 1\big) \\ &= \frac{1}{4^{40}}\big(91390 \times 81 + 9880 \times 27 + 780 \times 9 + 40 \times 1 + 1 \times 1\big) \\ &\simeq 6{,}35 \times 10^{-18}. \end{aligned}$$

Cette probabilité est infime. Elle est du même ordre de grandeur que celle de trouver 10 milliards de fois *consécutives* (sans tricher !) les 6 bons numéros au Loto en jouant une grille à 6 numéros à chaque tirage. Pour cela, l'heureux candidat et ses descendants devraient jouer pendant 100 millions d'années en gagnant à chaque fois ! On ne prend donc aucun risque en pratique en considérant qu'un candidat ayant obtenu au moins 36 bonnes réponses n'a

6. En fait dans cet exemple, le caractère unilatéral vient de ce que l'examinateur se demande si le candidat a fait *mieux* que de répondre au hasard. Et si le nombre de bonnes réponses obtenu est vraiment petit, par exemple $S = 0$, il convient de s'interroger sur le comportement du candidat. Est-il particulièrement malchanceux ? Ou est-ce un « rebelle » qui a fait exprès de répondre faux à chaque question ?

certainement pas répondu au hasard à *toutes* les questions. Néanmoins, il serait aventureux d'en conclure qu'un candidat ayant 37 bonnes réponses connaissait toutes ces réponses.

Code de la Route II. Dans un modèle plus réaliste, le candidat répond à coup sûr lorsqu'il connaît la réponse à la question et s'il l'ignore, choisit au hasard entre les 4 réponses proposées. On suppose que toutes les questions sont indépendantes et que pour chacune de ces questions, la probabilité que le candidat connaisse la vraie réponse est p. Ce paramètre p mesure donc le vrai niveau du candidat.

On peut alors vérifier (avec un peu de conditionnement) que :

- le nombre S de réponses connues du candidat suit la loi $\mathrm{Bin}(40, p)$;
- le nombre U de bonnes réponses du candidat suit la loi $\mathrm{Bin}\left(40; \dfrac{1+3p}{4}\right)$;
- le nombre T de bonnes réponses « chanceuses » données par le candidat suit la loi $\mathrm{Bin}\left(40; \dfrac{1-p}{4}\right)$.

Code de la Route III. Ce que peut réellement observer l'examinateur, c'est la valeur prise par U. Les quantités intéressantes pour tirer des conclusions sur le niveau réel du candidat sont les $P(S = i \mid U = m)$ pour $i \leqslant m \leqslant 40$. Par exemple si le candidat a obtenu 38 bonnes réponses, on aimerait évaluer $P(S \geqslant 36 \mid U = 38)\ldots$

Un corrigé détaillé du problème sur le Code de la Route est consultable à l'URL :
http://math.univ-lille1.fr/~suquet/Polys/CodeRouteCor.pdf

Revenons à nos histoires d'urnes. Dans le problème à deux urnes décrit ci-dessus, la séparation des masses entre les deux lois binomiales mises en concurrence (cf. figure 4) était bien nette. Par contre dans le problème suivant avec une seule urne, la situation est bien plus complexe. En effet, il s'agit maintenant de *mettre en concurrence* la loi binomiale $\mathrm{Bin}(100; 0{,}3)$ correspondant à la composition d'urne que l'on cherche à « tester » avec *toutes les autres lois binomiales*, $\mathrm{Bin}(100, p)$ pour $p \neq 0{,}3$ (pour p rationnel de $]0, 1[$ puisqu'il s'agit d'une proportion [7]).

Et là, on voit poindre une difficulté. Si p est trop « proche » de 0,3, les intervalles de fluctuation de $\mathrm{Bin}(100; 0{,}3)$ et de $\mathrm{Bin}(100, p)$ vont largement se recouvrir et la règle de décision proposée ci-dessus perdra beaucoup de sa pertinence. Alors que faire ?

La première réponse est d'augmenter le nombre de tirages n. En effet, en prenant n assez grand, on arrivera toujours à rendre disjoints les intervalles

7. Donc cela fait une infinité de lois concurrentes. En réalité, si on s'accorde sur un volume maximal d'urne, par exemple 1 m^3 et un diamètre minimal des boules, par exemple 1 mm, cela n'en fait plus qu'un nombre fini, mais quand même très grand.

de fluctuation au seuil de 95% de $\mathrm{Bin}(n;0{,}3)$ et de $\mathrm{Bin}(n,p)$. Pour $p < 0{,}3$, il suffit que $np + \sqrt{n} < 0{,}3n - \sqrt{n}$; pour $p > 0{,}3$, il suffit que $np - \sqrt{n} > 0{,}3n + \sqrt{n}$. Ces deux conditions peuvent se résumer par :

$$n > \frac{4}{(p-0{,}3)^2}.$$

Par exemple avec $p = 0{,}4$, il faudrait au moins 400 tirages pour avoir des intervalles de fluctuation disjoints, avec $p = 0{,}31$, il en faudrait 40 000.

La seconde réponse est que la première n'est pas réaliste ! En effet, en pratique, il faut bien se fixer un nombre de tirages et comme on ne connaît pas p, on ne saura jamais si on a une bonne séparation des intervalles de fluctuation, comme dans le problèmes des deux urnes ci-dessus. Autrement dit, il faut se faire à l'idée que lorsque l'on prétend valider l'hypothèse que la proportion inconnue est 0,3, la valeur 0,3 est donnée avec une certaine précision, dépendant du nombre d'observations que l'on peut se permettre et que des valeurs proches seraient aussi acceptables.

3 Estimation par intervalles de confiance

Pour présenter la notion d'intervalle de confiance, j'utiliserai le corrigé d'un exercice proposé récemment à mes étudiants du Master 1 MEFM (Métiers de l'Enseignement et de la Formation, en français moins obscur, « préparation au CAPES »).

Calibrage de pommes

Une coopérative agricole a un contrat de fourniture de pommes de catégorie A, c'est-à-dire dont le diamètre en mm est dans l'intervalle $[67, 73]$. Le gérant de la coopérative a besoin d'évaluer rapidement [8] la proportion p de pommes *hors catégorie* A dans la récolte qu'il vient d'emmagasiner. Pour cela, il prélève au hasard un échantillon de 400 pommes dont une calibreuse mécanique lui permet d'enregistrer les diamètres. On dénombre 70 pommes hors catégorie dans cet échantillon de taille 400. À partir de cette observation, nous allons construire deux intervalles de confiance au niveau 95% pour la proportion inconnue p de pommes hors catégorie dans la population totale.

Réduction à une situation binomiale. Remarquons d'abord que l'échantillon observé résulte d'un tirage *sans remise* de 400 individus dans une population de grande taille [9] N. En toute rigueur, la loi du nombre S_n de pommes hors

8. Avant de lancer l'opération de calibrage et d'emballage.

9. En comptant 6 pommes au kg, cela donnerait environ $N = 600\,000$ pommes pour une récolte de 100 tonnes.

catégorie dans l'échantillon de taille $n = 400$ est donc hypergéométrique de paramètres N, pN et n, mais vu la taille de la population, il est légitime d'approximer cette loi par la loi binomiale de paramètres $n = 400$ et p (inconnue) qui serait celle de S_n si les observations faites résultaient d'un tirage *avec remise*.

Dans toute la suite, nous supposerons donc que S_n suit la loi binomiale $\mathrm{Bin}(400, p)$.

Méthode avec variance majorée. En notant

$$S_n^* := \frac{S_n - np}{\sqrt{np(1-p)}} = \sqrt{\frac{n}{p(1-p)}}\left(\frac{S_n}{n} - p\right),$$

la somme centrée réduite, le théorème de de Moivre-Laplace nous dit que

$$\forall t > 0, \quad P(|S_n^*| \leqslant t) \xrightarrow[n\to+\infty]{} P(|Z| \leqslant t) = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1,$$

en utilisant pour la dernière égalité la symétrie de la loi de Z. On peut réécrire cette convergence sous la forme :

$$\forall n \geqslant 1, \forall t > 0, \quad P(-t \leqslant S_n^* \leqslant t) = 2\Phi(t) - 1 + \varepsilon_n(t), \quad \varepsilon_n(t) \xrightarrow[n\to+\infty]{} 0,$$

où $\varepsilon_n(t)$ est l'erreur d'approximation gaussienne.

En résolvant par rapport à p l'encadrement $-t \leqslant S_n^* \leqslant t$, on voit que

$$-t \leqslant S_n^* \leqslant t \quad \Longleftrightarrow \quad \frac{S_n}{n} - t\sqrt{\frac{p(1-p)}{n}} \leqslant p \leqslant \frac{S_n}{n} + t\sqrt{\frac{p(1-p)}{n}}.$$

Cet encadrement n'est pas satisfaisant pour construire un intervalle de confiance pour p car *les bornes dépendent de* p via la quantité inconnue $p(1-p)$ qui est la variance d'une v.a. de Bernoulli de paramètre p. La méthode avec variance majorée consiste à se débarasser de cet inconvénient en notant que la fonction $g : p \mapsto p(1-p)$ atteint son maximum sur $[0,1]$ au point $p_0 = \frac{1}{2}$ (figure 5) et est donc majorée par $g(p_0) = \frac{1}{4}$. Ainsi pour tout $p \in [0,1]$, $\sqrt{p(1-p)} \leqslant \sqrt{\frac{1}{4}} = \frac{1}{2}$, d'où :

$$-t \leqslant S_n^* \leqslant t \quad \Longrightarrow \quad \frac{S_n}{n} - \frac{t}{2\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t}{2\sqrt{n}}.$$

Cette implication se traduit par l'inclusion d'évènements $A_{n,t} \subset B_{n,t}$, en notant

$$A_{n,t} = \{-t \leqslant S_n^* \leqslant t\}, \quad B_{n,t} = \left\{\frac{S_n}{n} - \frac{t}{2\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t}{2\sqrt{n}}\right\}.$$

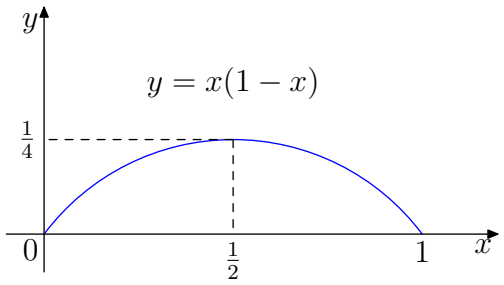

FIGURE 5 – Majoration de la variance d'une loi de Bernoulli

On en déduit que

$$P(B_{n,t}) \geqslant P(A_{n,t}) = 2\Phi(t) - 1 + \varepsilon_n(t).$$

On cherche t tel que $2\Phi(t) - 1 = 0{,}95$, c'est-à-dire $\Phi(t) = 0{,}975$, d'où $t = 1{,}96$ (par lecture inverse de la table des valeurs de Φ). Considérant alors que $n = 400$ est suffisamment grand pour que l'on puisse négliger l'erreur d'approximation gaussienne $\varepsilon_n(t)$, on obtient finalement $P(B_{n,t}) \geqslant 0{,}95$ pour $n = 400$ et $t = 1{,}96$. En introduisant l'intervalle *aléatoire* :

$$I_{n,t} = \left[\frac{S_n}{n} - \frac{t}{2\sqrt{n}}, \frac{S_n}{n} + \frac{t}{2\sqrt{n}}\right]$$

on peut réécrire ceci sous la forme

$$P(p \in I_{n,t}) \geqslant 0{,}95, \quad \text{pour } t = 1{,}96.$$

On pourra dire que $I_{n,t}$ est un intervalle de confiance *théorique* au niveau approximatif $2\Phi(t) - 1$, soit 95% pour $t = 1{,}96$.
Remarque : si on majore 1,96 par 2 que retrouve-t-on [10] ?

En réalité, ce que l'on a observé, c'est *une réalisation particulière* $S_n(\omega_0) = 70$ de la variable aléatoire S_n et comme on ne connaît pas p, on ne peut pas dire avec certitude si le ω_0 sous-jacent [11] appartient ou non à $B_{n,t}$. On « parie » donc sur la réalisation de $B_{n,t}$ et d'après l'étude précédente, la probabilité de gagner ce pari est (approximativement) au moins 95%. L'intervalle $I =$

10. Voir à ce sujet la discussion sur intervalle de fluctuation et intervalle de confiance, page 16.

11. Ici ω_0 représente les résultats observés de la suite de tirages effectués. Cela peut être une suite binaire si on code par 1 une pomme hors-catégorie et par 0 une pomme de catégorie A, ou une suite de réels si on enregistre vraiment le diamètre de chaque pomme prélevée, etc.

$I_{n,t}(\omega_0)$ est un intervalle de confiance *numérique*[12] pour p au niveau 95% :

$$I = \left[\frac{70}{400} - \frac{1{,}96}{2\sqrt{400}}; \frac{70}{400} + \frac{1{,}96}{2\sqrt{400}}\right] = [0{,}126; 0{,}224].$$

Remarque : attention à l'erreur classique « $P(p \in [0{,}126; 0{,}224]) = 0,95$ ». Quand on écrit « $p \in I_{n,t}$ », il s'agit de l'évènement $B_{n,t}$ auquel on peut attribuer une probabilité. Explicitement, $B_{n,t} = \{\omega \in \Omega;\ \frac{S_n(\omega)}{n} - \frac{t}{2\sqrt{n}} \leqslant p \leqslant \frac{S_n(\omega)}{n} + \frac{t}{2\sqrt{n}}\}$. Mais dès que l'on remplace $I_{n,t}$ par I, les bornes de l'intervalle ne dépendent plus de ω et comme p est inconnu mais pas aléatoire (p ne dépend pas de ω), l'ensemble $\{\omega \in \Omega;\ 0{,}126 \leqslant p \leqslant 0{,}224\}$ ne peut être que $\varnothing$ (si p ne vérifie pas l'encadrement) ou Ω (si p le vérifie). Si l'on considère cet ensemble comme un évènement, sa probabilité ne peut être que 0 ou 1, mais certainement pas 0,95.

Méthode avec variance estimée. Au lieu de majorer $p(1-p)$, on l'estime par $V_n = \frac{S_n}{n}(1 - \frac{S_n}{n})$. On vérifie facilement grâce à la loi forte des grands nombres que V_n converge presque sûrement, donc aussi en probabilité, vers $p(1-p)$. En notant

$$C_{n,t} = \left\{\frac{S_n}{n} - \frac{t\sqrt{V_n}}{\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t\sqrt{V_n}}{\sqrt{n}}\right\},$$

on obtient

$$P(C_{n,t}) = 2\Phi(t) - 1 + \varepsilon'_n(t),$$

puis avec le choix $t = 1{,}96$, $P(C_{n,t}) \simeq 0{,}95$. L'intervalle de confiance numérique correspondant est

$$J = [0{,}137; 0{,}213].$$

Contrairement aux apparences, ce n'est pas de la cuisine, à cause du

Théorème 3 (TLC avec autonormalisation)**.** *Soient $X_1, \ldots, X_n, \ldots$ des variables aléatoires indépendantes et de même loi telle que $\mathbf{E}\, X_1^2 < +\infty$ et $\sigma^2 := \operatorname{Var} X_1 > 0$. On note $S_n := X_1 + \cdots + X_n$. On suppose de plus que $(V_n)_{n\geqslant 1}$ est une suite de variables aléatoires positives qui converge en probabilité vers σ^2. Alors*

$$T_n := \sqrt{\frac{n}{V_n}}\left(\frac{S_n}{n} - \mathbf{E}\, X_1\right) \xrightarrow[n\to+\infty]{\text{loi}} Z,$$

où Z suit la loi gaussienne standard $\mathfrak{N}(0,1)$.

12. Les appelations intervalle de confiance *théorique* ou *numérique* ne sont pas standard et la plupart du temps dans la littérature statistique, on n'explicite pas la distinction, le contexte permettant à un lecteur un peu familier du sujet de lever l'ambiguïté.

En général on applique ce théorème en prenant pour V_n la « variance empirique ». Cette variance empirique est pour chaque ω, la variance calculée sur la série statistique réellement observée $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$. C'est donc la variable aléatoire :

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2,$$

où $\overline{X} = S_n/n$. En appliquant deux fois la loi forte des grands nombres, on voit facilement que la variance empirique converge presque sûrement (donc aussi en probabilité) vers la variance théorique $\sigma^2 = \mathbf{E}\, X_1^2 - (\mathbf{E}\, X_1)^2$.

Si on revient au cas où les X_i sont des variables de Bernoulli, on peut appliquer directement le théorème ci-dessus avec $V_n = \overline{X}(1 - \overline{X})$ en notant que par la loi forte des grands nombres, V_n converge presque-sûrement vers $p(1-p) = \sigma^2$. On peut aussi remarquer que pour des variables de Bernoulli, la variance empirique n'est autre que $\overline{X}(1 - \overline{X})$. La vérification de cette affirmation est facile une fois que l'on a noté que si X_i ne prend que les valeurs 0 et 1, $X_i = X_i^2$.

Un T.P. sur les intervalles de confiance.

Pour illustrer informatiquement les intervalles de confiance calculés par les deux méthodes, on peut utiliser un script Scilab téléchargeable à :
http://math.univ-lille1.fr/~suquet/Prog/peigne2-2011.sce

Ce script peut se décomposer en 3 sous-programmes décrits ci-dessous indépendamment de l'utilisation de tel ou tel langage de programmation [13].

Le sous-programme 1 demande à l'utilisateur une taille d'échantillon n et une valeur de probabilité p à estimer. Il génére ensuite 100 échantillons de taille n de la loi de Bernoulli de paramètre p. Il fournit ensuite ces 100 échantillons au sous-programme 2.

Le sous-programme 2 ignore la valeur de p, il ne connaît que n et les échantillons générés ci-dessus. Pour chacun de ces 100 échantillons, il calcule les deux intervalles de confiance (méthode avec variance majorée et méthode avec variance estimée).

Le sous-programme 3 connaît p et regarde pour chacun des intervalles de confiance calculés s'il contient p (dans ce cas il colorie l'intervalle en vert) ou non (dans ce cas il colorie l'intervalle en rouge). Pour chacune des deux méthodes, il trace ensuite en bleu le segment horizontal d'équation $y = p$, $x \in [0, 100]$ et les intervalles de confiance représentés à la verticale de leur numéro (de 1 à 100). Enfin, il imprime à l'écran les numéros des échantillons pour lesquels l'intervalle de confiance rate p.

13. Dans le script `peigne2-2011.sce`, la séparation en 3 sous-programmes n'apparaît pas aussi clairement que dans la description qui suit car j'ai voulu économiser une boucle.

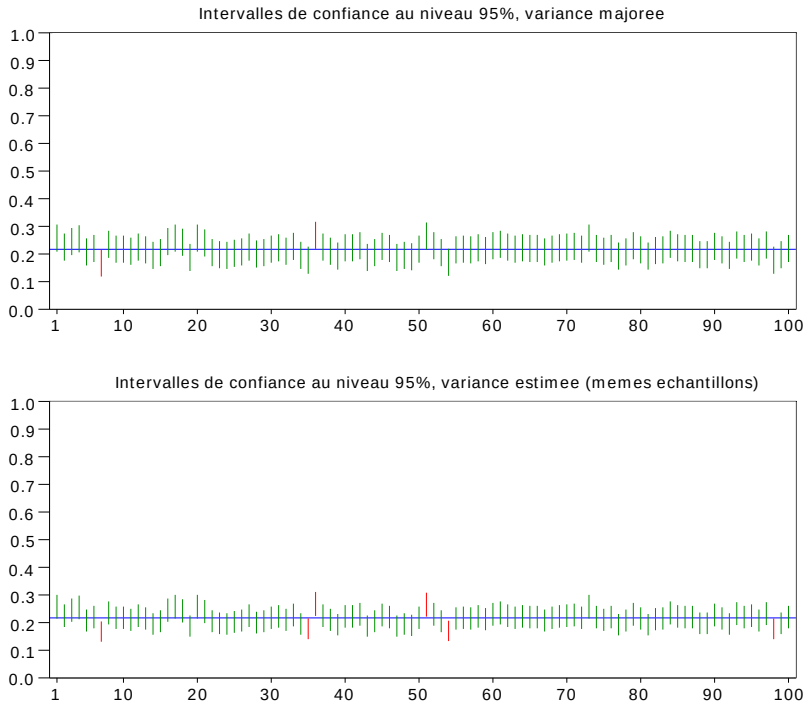

FIGURE 6 – Peignes d'intervalles de confiance, 100 échantillons de taille 400, $p = 0{,}217$

La figure 6 représente le résultat d'une exécution du script, avec une taille d'échantillon 400 et $p = 0{,}217$. Les échantillons dont l'intervalle de confiance avec variance majorée rate p sont les n° 7 et 36. Ceux dont l'intervalle de confiance avec variance estimée rate p sont les n° 7, 35, 36, 51, 54, 98. Il n'est pas surprenant de trouver plus d'intervalles ratant p pour la méthode avec variance estimée car ces intervalles sont plus courts [14].

Le sous-programme 1 représente « l'état de la nature », et la collecte de données que l'on est bien obligé de simuler ici. Le sous-programme 2 représente le statisticien qui veut estimer la quantité inconnue p à partir des données collectées. Dans la « vraie vie », le sous-programme 3 n'existe pas ou sa réalisation est souvent trop difficile, trop coûteuse ou inacceptable [15].

Fluctuation ou confiance [16] *?*

Lorsque l'on a calculé l'intervalle de confiance théorique par la méthode avec variance majorée, on a obtenu un intervalle de bornes $S_n/n \pm 1{,}96/(2\sqrt{n})$. Et la probabilité que cet intervalle aléatoire contienne p est approximativement 0,95. En majorant 1,96 par 2, on élargit l'intervalle, ce qui ne diminue pas cette probabilité approximative de 0,95. L'intervalle ainsi obtenu s'écrit alors :

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

et on peut aussi évaluer la probabilité que cet intervalle contienne p à environ 0,95 (plus précisément (sic), cette probabilité est supérieure ou égale à la précédente que l'on approximait déjà par 0,95).

Le nouvel intervalle aléatoire obtenu ci-dessus fait fortement penser à la formule d'approximation pour l'intervalle de fluctuation. On aurait d'ailleurs pu obtenir directement ce nouvel intervalle de confiance théorique élargi en utilisant l'approximation de l'intervalle de fluctuation. En effet, pour toutes valeurs de n et p pour lesquelles cette approximation est valide, on a

$$0{,}95 \simeq P\left(p - \frac{1}{\sqrt{n}} \leqslant \frac{S_n}{n} \leqslant p + \frac{1}{\sqrt{n}}\right)$$

14. Comme toujours en statistique, on ne peut pas gagner sur tous les tableaux. Les intervalles avec variance estimée donnent un encadrement plus précis pour p, au prix d'un niveau de confiance moins élevé que pour la méthode avec variance majorée. L'élargissement d'intervalle dû à la majoration de la variance augmente en général la probabilité que cet intervalle contienne p, ce qui peut donner un niveau de confiance exact (mais pas toujours calculable) supérieur à 95%.

15. Si on veut connaître avec exactitude la proportion p de carpes parmi les poissons d'un étang, la seule méthode sûre consiste à vider l'étang.

16. Les remarques qui suivent sur la comparaison entre intervalle de fluctuation et intervalle de confiance n'ont pas été développées lors de l'exposé. C'était certainement un manque.

et comme l'encadrement ci-dessus de S_n/n se résout en un encadrement de p, on en déduit facilement que :

$$P\left(\frac{S_n}{n} - \frac{1}{\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{1}{\sqrt{n}}\right) \simeq 0{,}95.$$

Ceci pourrait laisser croire qu'intervalle de fluctuation et intervalle de confiance, sont la même chose. Mais il n'en est rien et il est utile de clarifier cette question. Les deux intervalles ne sont pas de même nature :

- l'intervalle de fluctuation $[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}]$ a des bornes *déterministes* et dépendantes de p qui est ici un paramètre inconnu. On ne sait en général pas les calculer (sauf dans le problème de prise de décision discuté à la section 2).
- l'intervalle de confiance théorique $[\frac{S_n}{n} - \frac{1}{\sqrt{n}}, \frac{S_n}{n} + \frac{1}{\sqrt{n}}]$ a des bornes *aléatoires* et on peut en calculer une réalisation (intervalle de confiance numérique) au vu des observations.

Fluctuation ou confiance, suite à sauter en première lecture[17].

Si on veut approfondir un peu la question, il faut en passer par la notion de modèle statistique et donner une définition formelle de la notion d'intervalle de confiance théorique. On verra que comme souvent, le diable est dans les notations et que les écritures utilisées ci-dessus comme « $P(p \in I_{n,t})$ » ou « $P(\frac{S_n}{n} \in [p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}])$ » ont une simplicité trompeuse.

Pour simplifier, nous travaillerons avec n fixé. L'univers le plus simple pouvant représenter les issues élémentaires de l'expérience de calibrage des pommes (assimilée à un tirage avec remise) est l'ensemble $\Omega_n = \{0,1\}^n$ des suites binaires $\omega = (u_1, \ldots, u_n)$ de longueur n, en codant $u_i = 0$ pour une pomme de catégorie A au i^{e} tirage et $u_i = 1$ pour une pomme hors catégorie. Comme Ω_n est fini, on prendra comme famille $\mathcal{F}_n$ d'évènements observables, la famille de tous les sous-ensembles de Ω. À ce stade, on peut déjà définir les variables aléatoires X_i et S_n en posant :

$$\forall \omega = (u_1, \ldots, u_n), \quad X_i(\omega) = u_i, \quad S_n(\omega) = \sum_{i=1}^{n} X_i(\omega) = \sum_{i=1}^{n} u_i.$$

Remarquons que la définition de ces variables aléatoires ne fait intervenir *aucune notion de probabilité* sur l'espace $(\Omega_n, \mathcal{F}_n)$. En particulier, ces variables ne dépendent pas de p.

Maintenant se pose la question : de quelle probabilité P allons nous munir $(\Omega_n, \mathcal{F}_n)$? Si on connaît la proportion p de pommes hors catégorie A dans la récolte, la réponse est donnée par la formule :

$$\forall \omega = (u_1, \ldots, u_n) \in \Omega_n, \quad P(\{\omega\}) = p^{S_n(\omega)}(1-p)^{n - S_n(\omega)}$$

17. Et probablement aussi en deuxième.

qui traduit l'indépendance des tirages avec remise[18]. Comme Ω_n est fini, il suffit de donner la probabilité de chaque évènement élémentaire ω pour pouvoir définir de manière unique la probabilité d'un évènement quelconque B par $P(B) = \sum_{\omega \in B} P(\{\omega\})$.

Mais dans le problème d'estimation de p, on ne connaît pas p. Il faut donc se résigner à munir $(\Omega_n, \mathcal{F}_n)$ non pas d'*une* probabilité P, mais de *toute une famille* $(P_p)_{p \in]0,1[}$ de probabilités. Pour chaque valeur de p, P_p est définie comme ci-dessus par $P_p(\{\omega\}) = p^{S_n(\omega)}(1-p)^{n-S_n(\omega)}$. On obtient ce que l'on appelle un *modèle statistique* :

$$\Big(\Omega_n, \mathcal{F}_n, (P_p)_{p \in]0,1[}\Big).$$

Nous avons déjà noté que les variables aléatoires X_i et S_n ne dépendent pas de p. Il n'en va pas de même pour leur *loi*. Mais ici il convient de rappeler que la loi d'une variable aléatoire Y définie sur un espace $(\Omega, \mathcal{F})$ n'est pas une propriété intrinsèque de la variable aléatoire. Elle dépend de la probabilité P dont on munit $(\Omega, \mathcal{F})$ et est caractérisée par les $P(Y \in I)$, pour I intervalle[19] de $\mathbb{R}$. Au lieu de parler de la loi de Y, il serait donc plus correct de parler de la loi de Y *sous* P. Dans les problèmes de probabilité, l'espace $(\Omega, \mathcal{F})$ est muni d'une seule probabilité P et ce distinguo est superflu. Par contre, quand on travaille avec un modèle statistique comme ci-dessus, le distinguo devient crucial et il faut parler de la loi de Y sous P_p. Notons en passant que la notion d'indépendance d'évènements ou de variables aléatoires n'est pas davantage intrinsèque et qu'elle est, elle aussi, relative à la probabilité P dont on a muni l'espace $(\Omega, \mathcal{F})$.

Dans le cadre de notre modèle statistique $\big(\Omega_n, \mathcal{F}_n, (P_p)_{p \in]0,1[}\big)$, nous pouvons noter ce qui suit.

a) *Pour chaque* $p \in]0,1[$, les X_i sont P_p-indépendantes et de même loi sous P_p, à savoir la loi de Bernoulli de paramètre p.

18. Vous trouvez peut-être surprenant d'utiliser une variable aléatoire S_n pour définir une probabilité, mais ici $S_n(\omega)$ n'est rien d'autre que le nombre de 1 dans la suite binaire $\omega = (u_1, \ldots, u_n)$ ou nombre de succès obtenus au cours des n tirages que réprésente ω et $n - S_n(\omega)$ est le nombre de zéros ou nombre d'échecs.

19. On peut se demander pourquoi je parle de variable aléatoire définie sur $(\Omega, \mathcal{F})$ et pas sur Ω. C'est précisément parce que P est définie sur la famille d'évènements $\mathcal{F}$ et que pour pouvoir donner un sens aux quantités $P(Y \in I)$, il faut que les ensembles $\{Y \in I\} = \{\omega \in \Omega; Y(\omega) \in I\}$ appartiennent à cette famille $\mathcal{F}$. En fait on appelle variable aléatoire définie sur $(\Omega, \mathcal{F})$, toute application de Ω dans $\mathbb{R}$ vérifiant cette propriété (c'est la *mesurabilité*). Bien sûr, quand $\mathcal{F}$ est la famille de toutes les parties de Ω, cette condition de mesurabiltié est automatiquement vérifiée. Mais si on veut modéliser une suite infinie de tirages avec remise, on prendra pour Ω l'ensemble des suites binaires infinies et là, on ne peut plus prendre pour $\mathcal{F}$ la famille des toutes les parties de Ω et obtenir une modélisation compatible avec celle des n tirages.

b) *Pour chaque* $p \in]0,1[$, la loi de S_n sous P_p est la binomiale de paramètres n et p.

Nous dirons qu'une suite de variables aléatoires $X_1, \ldots, X_n$ sur $(\Omega_n, \mathcal{F}_n)$, pas forcément définies comme ci-dessus, est un *échantillon* associé au modèle si elle vérifie la propriété a) ci-dessus.

De même, les théorèmes limite (loi des grands nombres et théorèmes limite centraux) devraient être réécrits en remplaçant convergence presque sûre par convergence P_p-p.s. et convergence en loi par convergence en P_p-loi, pour tout $p \in]0,1[$.

Nous pouvons maintenant formaliser une définition mathématique des intervalles de confiance théoriques [20].

Définition 4. *Soient* $\big(\Omega_n, \mathcal{F}_n, (P_p)_{p\in]0,1[}\big)$ *un modèle statistique,* $X_1, \ldots, X_n$ *un échantillon associé au modèle et* $\varepsilon \in]0,1[$. *On appelle intervalle de confiance théorique pour* p *de niveau au moins* $1-\varepsilon$, *tout intervalle fermé dont les bornes sont des variables aléatoires* $a_n(X_1, \ldots, X_n)$ *et* $b_n(X_1, \ldots, X_n)$ *vérifiant :*

$$\inf_{p\in]0,1[} P_p\big(a_n(X_1, \ldots, X_n) \leqslant p \leqslant b_n(X_1, \ldots, X_n)\big) \geqslant 1-\varepsilon.$$

Cette condition aurait pu aussi s'écrire en remplaçant « $\inf_{p\in]0,1[}$ » par « $\forall p \in]0,1[$ ». Ce ne serait plus vrai si on remplaçait la dernière inégalité par une égalité à $1-\varepsilon$. Il convient de noter ici que la fonction $\inf_{p\in]0,1[} P_p$ définie sur $\mathcal{F}_n$ par $A \mapsto \inf_{p\in]0,1[} P_p(A)$ n'est en général pas une probabilité. C'est pour cela que l'on dit que l'intervalle $[a_n, b_n]$ a un *niveau de confiance* d'au moins $1-\varepsilon$ (pour l'encadrement de p) et pas une probabilité d'au moins $1-\varepsilon$.

Le point essentiel dans la définition de l'intervalle de confiance est que dans les inégalités de minoration :

$$\forall p \in]0,1[, \quad P_p\big(a_n(X_1, \ldots, X_n) \leqslant p \leqslant b_n(X_1, \ldots, X_n)\big) \geqslant 1-\varepsilon,$$

le minorant $1-\varepsilon$ *ne dépend pas de* p. Il s'agit donc d'une minoration *uniforme* sur la famille $(P_p)_{p\in]0,1[}$. En fait, pour n fixé, on se contentera plus modestement en pratique d'une minoration uniforme sur une sous-famille $(P_p)_{p\in[\delta_n,1-\delta_n]}$, par exemple en raison de la dégradation de l'approximation de la loi binomiale par une gaussienne lorsque p est proche de 0 ou de 1. Le théorème de de Moivre-Laplace, comme la formule d'approximation de l'intervalle de fluctuation au seuil de 95% sont des outils permettant d'obtenir

20. En réalité, on peut proposer plusieurs définitions, en distinguant niveau au moins $1-\varepsilon$ ou niveau exact $1-\varepsilon$, intervalle de confiance asymptotique ou non, etc. La définition retenue ici est un compromis suffisant pour nos besoins.

une telle minoration uniforme. Pour l'intervalle de fluctuation approché, on peut ainsi écrire que :

$$\forall n \geqslant 25,\ \forall p \in [0{,}2; 0{,}8], \quad P_p\Big(\frac{S_n}{n} \in \Big[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\Big]\Big) \geqslant 0{,}95.$$

4 Contrôle de l'erreur d'approximation gaussienne

Dans l'approximation d'une loi binomiale par une gaussienne, on ne peut espérer en général avoir une précision d'un ordre meilleur que $O(n^{-1/2})$.

Voici un exemple élémentaire avec S_{2n} de loi $\mathrm{Bin}(2n, \frac{1}{2})$, d'où $\mathbf{E}\, S_{2n} = n$. On cherche un équivalent de $P(S^*_{2n} \leqslant 0) - \Phi(0)$. Remarquons d'abord que

$$\{S^*_{2n} < 0\} = \{0 \leqslant S_{2n} < n\} \quad \text{et} \quad \{S^*_{2n} > 0\} = \{n < S_{2n} \leqslant 2n\}.$$

En raison de la symétrie des coefficients binomiaux ($C^k_{2n} = C^{2n-k}_{2n}$),

$$P(S^*_{2n} < 0) = \sum_{k=0}^{n-1} C^k_{2n} 2^{-2n} = \sum_{j=n+1}^{2n} C^j_{2n} 2^{-2n} = P(S^*_{2n} > 0).$$

On a ainsi $2P(S^*_{2n} < 0) + P(S^*_{2n} = 0) = 1$ d'où l'on tire :

$$P(S^*_{2n} < 0) = \frac{1}{2} - \frac{1}{2} P(S^*_{2n} = 0), \quad P(S^*_{2n} \leqslant 0) = \frac{1}{2} + \frac{1}{2} P(S^*_{2n} = 0).$$

En rappelant que $\Phi(0) = \frac{1}{2}$, on aboutit à :

$$P(S^*_{2n} \leqslant 0) - \Phi(0) = \frac{1}{2} P(S^*_{2n} = 0) = \frac{1}{2} P(S_{2n} = n) = C^n_{2n} 2^{-2n-1}.$$

Par la formule de Stirling ($n! \sim \sqrt{2\pi} n^{n+1/2} \mathrm{e}^{-n}$), on obtient l'équivalent

$$P(S^*_{2n} \leqslant 0) - \Phi(0) \sim \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2n}}.$$

Comme $(2\pi)^{-1/2} > 0{,}398\,9$, on a $|P(S^*_{2n} \leqslant 0) - \Phi(0)| \geqslant 0{,}398 (2n)^{-1/2}$, pour $n \geqslant n_0$.

Voici maintenant un théorème général sur la vitesse de convergence dans le théorème limite central.

Théorème 5 (Berry-Esséen, 1941–42)**.** *Soit $(X_i)_{i \geqslant 1}$ une suite de variables aléatoires indépendantes et de même loi, telle que $\mathbf{E}\,|X_i|^3 < +\infty$. On note*

$\sigma^2 := \operatorname{Var} X_1$, $\rho^3 := \mathbf{E}\,|X_1 - \mathbf{E}\,X_1|^3$, *avec* $\sigma > 0$ *et* $\rho > 0$. *Il existe alors une constante universelle* $C > 0$ *telle que pour tout* $n \geqslant 1$,

$$\Delta_n := \sup_{x\in\mathbb{R}} \big|P(S_n^* \leqslant x) - \Phi(x)\big| \leqslant C\frac{\rho^3}{\sigma^3}\,\frac{1}{\sqrt{n}}.$$

On pourra trouver une preuve du théorème de Berry-Esséen dans le tome 2 du célèbre ouvrage de Feller [1]. L'obtention de la meilleure constante C a été l'objet d'une longue quête. La valeur initiale de Esséen était $C = 7{,}59$. Une valeur plus moderne et proche de l'optimale est $C = 0{,}797\,5$ (Van Beek (1972)).

Dans le cas de de Moivre-Laplace, donc avec des X_i suivant la loi de Bernoulli de paramètre p. On trouve

$$\Delta_n \leqslant C\frac{p^2+q^2}{\sqrt{pq}}\,\frac{1}{\sqrt{n}}, \quad q := 1-p.$$

Le coefficient $(p^2+q^2)(pq)^{-1/2}$ est minimal pour $p = q = 1/2$, et explose quand p tend vers 0 ou vers 1, ce qui est conforme aux illustrations graphiques qui permettent de voir que pour une valeur de n fixée, l'approximation gaussienne de la binomiale est la meilleure pour $p = 1/2$ et se dégrade sensiblement lorsque p est proche de 0 ou de 1.

Ceci dit, la borne obtenue via le théorème de Berry-Esséen vient d'un théorème universel, valable pour toutes les v.a. X_1 telles que $\mathbf{E}\,|X_1|^3 < +\infty$. On peut donc espérer l'améliorer en exploitant les spécificités de la loi binomiale. Comme nous l'avons vu ci-dessus, cette amélioration ne peut porter sur le facteur $n^{-1/2}$, mais seulement sur son coefficient. Ces résultats spécifiques à la loi binomiale sont dus à Uspensky [4]. Pour des énoncés précis en français, on pourra consulter le chapitre 7 de [2]. Voici un corollaire facile à mémoriser de ces résultats d'Uspensky.

Si $npq \geqslant 25$, alors pour tous réels $x_1 < x_2$,

$$\big|P(x_1 \leqslant S_n^* \leqslant x_2) - \big(\Phi(x_2) - \Phi(x_1)\big)\big| \leqslant \frac{0{,}588}{\sqrt{npq}}.$$

Ce majorant est effectivement meilleur que celui fourni par application brutale du théorème de Berry-Esséen, à savoir $2(p^2+q^2)(npq)^{-1/2}$. En effet, la quantité $2(p^2+q^2)$ a pour minimum 1, atteint pour $p = q = 1/2$ et pour maximum 2 atteint pour $p = 0$ ou 1. Il convient néanmoins de remarquer que la borne de Berry-Esséen est valide pour tout $n \geqslant 1$ et tout $p \in]0,1[$, tandis que la condition $npq \geqslant 25$ impose des valeurs de n assez grandes (au moins 100 dans le cas le plus favorable où $p = 1/2$).

5 À propos de l'intervalle de fluctuation

Cette partie n'a pas du tout été abordée lors de l'exposé. Elle contient quelques remarques critiques *a posteriori* sur la définition de l'intervalle de fluctuation.

Reprenons la définition : « l'intervalle de fluctuation au seuil de 95%, relatif aux échantillons de taille n, est l'intervalle *centré* autour de p, proportion du caractère dans la population, où se situe, avec une probabilité *égale* à 0,95, la fréquence observée dans un échantillon de taille n ».

Autrement dit, si S_n suit la loi binomiale de paramètres n et p, l'intervalle de fluctuation de la fréquence S_n/n est l'intervalle de la forme $[p-r, p+r]$ tel que

$$P\left(p - r \leqslant \frac{S_n}{n} \leqslant p + r\right) = 0{,}95,$$

ou de manière équivalente, l'intervalle de fluctuation de S_n est l'intervalle de la forme $[np - nr, np + nr]$ tel que

$$P(np - nr \leqslant S_n \leqslant np + nr) = 0{,}95.$$

Par la suite, pour des raisons de confort d'écriture, je ne parlerai que de l'intervalle de fluctuation pour S_n (le passage à l'intervalle de fluctuation pour la fréquence s'en déduisant immédiatement en divisant les bornes par n). Comme cet intervalle ne dépend *que de la loi* de S_n, on pourra parler d'intervalle de fluctuation de la loi binomiale $\mathrm{Bin}(n, p)$.

La définition de l'intervalle de fluctuation pose immédiatement deux questions :

1. existe-t-il toujours un tel intervalle ?
2. s'il existe, est-il unique ?

Malheureusement, la réponse à ces deux questions est non.

Pour la première question, il suffit de remarquer que la famille des probabilités $H = \{P(x \leqslant S_n \leqslant y);\ x, y \in \mathbb{R}\}$ est un sous-ensemble *fini* de $[0, 1]$. En effet la fonction de répartition de S_n est en escaliers et comporte $n + 2$ marches en tout, il n'y a donc qu'un nombre fini de dénivelés possibles entre deux marches quelconques. On ne voit vraiment pas pourquoi la valeur 0,95 figurerait systématiquement dans H.

Pour la deuxième question, supposons qu'on ait trouvé un réel $r > 0$ tel que $P(np - nr \leqslant S_n \leqslant np + nr) = 0{,}95$. Remarquons alors que la fonction de deux variables : $(x, y) \mapsto P(x \leqslant S_n \leqslant y)$ est constante sur les carrés de la forme $]j - 1, j] \times [k, k + 1[$, où j et k sont entiers. Il est clair alors qu'en augmentant ou diminuant légèrement r on peut trouver une infinité de solutions au problème.

Pour surmonter ces deux inconvénients de la définition de l'intervalle de fluctuation, on pourrait modifier la définition en disant qu'il s'agit du *plus court* intervalle fermé I centré sur p tel que $P(S_n/n \in I) \geqslant 0{,}95$.

Et finalement, on pourrait aussi bien s'intéresser au plus court intervalle fermé contenant S_n avec une probabilité d'au moins 95%. Cet intervalle pourrait être dénommé *intervalle de concentration* de S_n au seuil de 95% (en divisant ses bornes par n, on obtient évidemment l'intervalle de concentration pour la fréquence S_n/n). Il ne sera pas en général centré sur np, car la répartition des probabilités $P(S_n = k)$ n'est symétrique que dans le cas $p = 1/2$ et est d'autant plus asymétrique (pour n fixé) que p est éloigné de $1/2$. D'autre part, d'après la remarque ci-dessus sur $P(x \leqslant S_n \leqslant y)$, il est clair que les bornes de l'intervalle de concentration seront des entiers.

Il y a un algorithme simple pour trouver l'intervalle de concentration. Cet algorithme repose sur le lemme suivant qui nous décrit le sens de variation de la suite finie des $P(S_n = k)$, pour S_n de loi $\mathrm{Bin}(n, p)$. Dans ce qui suit, on note $[x]$ la partie entière[21] (inférieure) d'un réel x, autrement dit $[x]$ est l'unique entier m tel que $m \leqslant x < m + 1$.

Lemme 6. *Pour $p \in]0, 1[$ et $n \in \mathbb{N}^*$ fixés, posons $q = 1 - p$ et*

$$b_k = C_n^k p^q q^{n-k} = P(S_n = k), \quad \text{où} \quad S_n \sim \mathrm{Bin}(n, p).$$

Notons $m = [(n + 1)p]$ la partie entière de $(n + 1)p$. Alors la suite finie $(b_k)_{0 \leqslant k \leqslant n}$ a les variations suivantes.

1. *Si $(n+1)p$ n'est pas entier, la suite croît strictement sur $[\![0, m]\!]$ et décroît strictement sur $[\![m, n]\!]$. Elle a donc un unique maximum en $k = m$.*
2. *Si $(n + 1)p$ est entier, la suite croît strictement sur $[\![0, m - 1]\!]$, décroît strictement sur $[\![m, n]\!]$ et $b_{m-1} = b_m$. Le maximum est donc atteint en $m - 1$ et m.*

Preuve. Pour $k \in [\![1, n]\!]$, on calcule le rapport b_k/b_{k-1} et on le compare à 1.

$$\begin{aligned}\frac{b_k}{b_{k-1}} &= \frac{n!}{k!(n-k)!} \times \frac{(k-1)!(n-k+1)!}{n!} \times \frac{p^k q^{n-k}}{p^{k-1} q^{n-k+1}} \\ &= \frac{(n-k+1)p}{kq} = \frac{(n+1)p - kp}{kq} = \frac{(n+1)p - k(1-q)}{kq} \\ &= \frac{kq + (n+1)p - k}{kq}.\end{aligned}$$

21. La notation $\mathrm{E}(x)$ pour la partie entière prêterait évidement à confusion dans un contexte probabiliste.

On voit ainsi que :

$$\forall k \in [\![1, n]\!], \quad \frac{b_k}{b_{k-1}} = 1 + \frac{(n+1)p - k}{kq}.$$

Par conséquent, pour $k < (n+1)p$, $b_k/b_{k-1} > 1$, pour $k > (n+1)p$, $b_k/b_{k-1} < 1$ et l'égalité $b_k = b_{k-1}$ n'est possible que si $k = (n+1)p$, ce qui ne peut arriver que si $(n+1)p$ est entier. Le lemme est prouvé. □

Voici maintenant l'algorithme pour trouver l'intervalle de concentration.

1. Générer la liste B des probabilités binomiales $b_k = P(S_n = k)$, $k \in [\![0, n]\!]$.
2. La trier par valeurs décroissantes, on obtient ainsi une liste BT en mémorisant la liste L des indices initiaux.
3. Initialiser la somme s à $b_m = B(0)$ et la liste I à $I(0) = m$.
4. Parcourir BT en ajoutant chaque élement à s et en concaténant son indice initial (lu dans L) dans I, en s'arrêtant dès que $s \geqslant 0{,}95$.
5. Les bornes de l'intervalle de concentration sont alors le plus petit et le plus grand entier figurant dans I.

Remarquons que d'après les variations de la suite (b_k), cet algorithme ne laisse aucun « trou » dans I et que si on trie I dans l'ordre croissant, on obtiendra bien un intervalle d'entiers $[\![a, b]\!]$.

Évidemment, il y a quelques adaptations à faire en fonction du langage de programmation utilisé. Par exemple en Scilab, on ne peut pas indexer les vecteurs à partir de 0, il faut commencer à 1.

L'algorithme pour trouver l'intervalle de fluctuation est plus simple car il évite l'étape de tri. Si np n'est pas entier, on cumule successivement les deux plus proches voisins à gauche et à droite qui n'ont pas encore été pris dans la liste des b_k (en commençant donc avec $b_{[np]}$ et $b_{[np]+1}$), jusqu'à l'atteinte ou au franchissement du seuil 0,95. Dans ce cas, l'intervalle comportera un nombre pair d'entiers. Dans le cas particulier où np est entier, on commence avec b_{np} seul et on lui adjoint ses plus proches voisins gauche et droite non encore pris à chaque étape. L'intervalle de fluctuation obtenu aura alors un nombre impair d'entiers.

On voit bien ici que la contrainte de symétrie peut nous amener à un intervalle de fluctuation strictement plus grand que l'intervalle de concentration. Par exemple pour $\mathrm{Bin}(100; 0{,}3)$, en utilisant les algorithmes ci-dessus, voir [3] pour une implémentation en Scilab, on trouve que

– l'intervalle de concentration au seuil de 95% est $[\![22, 39]\!]$, avec une masse totale de 0,950 180 ;

– l'intervalle de fluctuation au seuil de 95% est $[\![21, 39]\!]$ avec une masse totale de 0,962 549.

Références

[1] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II. Wiley.

[2] Ch. Suquet, Introduction au Calcul des Probabilités, polycopié de L2 Lille 1. http://math.univ-lille1.fr/~suquet/Polys/ICP.pdf

[3] Ch. Suquet, bibliothèque de fonctions Scilab sur les lois binomiales, utilisée dans cet exposé.
http://math.univ-lille1.fr/~suquet/Prog/Bib.sci

[4] J. V. Uspensky, *Introduction to mathematical probability*. McGraw-Hill, 1937.