Introduction à la statistique inférentielle

Pierre Grihon(*)

« L'ignorance des différentes causes qui concourent à la production des événements, et leur complication, jointe à l'imperfection de l'Analyse, empêchent l'Homme de se prononcer avec la même certitude sur le plus grand nombre de phénomènes ; il y a donc pour lui des choses incertaines et d'autres plus ou moins probables. Dans l'impossibilité de les connaître, il a cherché à s'en dédommager en déterminant leurs différents degrés de vraisemblance, en sorte que nous devons à la faiblesse de l'esprit humain une des théories les plus délicates et les plus ingénieuses des mathématiques, savoir la science des hasards ou des probabilités. »(1)

Cet article a pour but :

- de mettre en perspective les nouvelles notions de statistique inférentielle introduites dans les programmes du lycée,
- de mettre en évidence les points essentiels et de préciser les enjeux de leur enseignement,
- de préciser la spécificité de leur approche en terminale.

Pour des précisions sur les notions figurant au programme de terminale, on pourra se reporter au document ressource publié sur Eduscol :
http://media.eduscol.education.fr/file/Mathematiques/11/5/LyceeGT_ressources_Math_T_proba-stat_207115.pdf

Pour commencer, voici un exemple que tout le monde connaît !

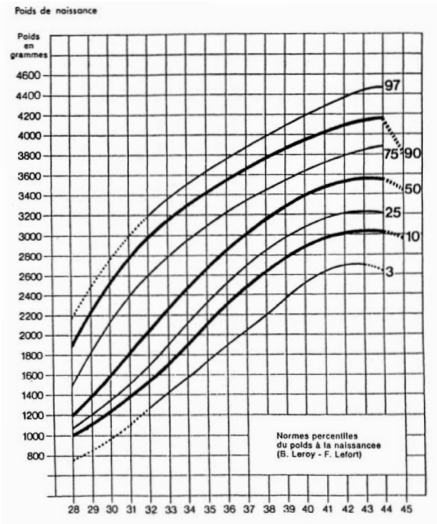
Poids des nouveau-nés

Les célèbres courbes de la page suivante, ainsi que celles figurant dans le carnet de santé de tous les enfants sont obtenues par des méthodes de statistiques inférentielles : elles donnent un encadrement du poids en gramme d'un nouveau-né en fonction de la durée de la grossesse. Pour les obtenir, on fait d'abord l'hypothèse que pour chaque durée de grossesse ce poids suit une loi normale de paramètres à déterminer, lesquels dépendent bien sûr de la durée de la grossesse.

Après observation du poids de naissance d'un certain nombre de nouveau-nés, on fait une *estimation* de la moyenne et de l'écart type du poids des futurs bébés. On peut alors calculer la probabilité que le poids des enfants nés à terme (40 semaines environ) se situe entre 2500g et 4200g : environ 0,94.

(*) pgrihon@free.fr

(1) Pierre-Simon de Laplace, Œuvres complètes, tome 8.



L'inférence statistique est un ensemble de méthodes permettant de tirer des conclusions sur une population à partir de données d'échantillons statistiques issus de cette population avec des informations quantifiées sur la fiabilité de ces conclusions.

Avant de traiter de statistiques inférentielles, il est nécessaire de faire quelques rappels de résultats de probabilités. Ceux-ci donnent les éléments théoriques nécessaires pour mettre en place les outils de statistique inférentielle dans le cadre binomial, le seul possible au niveau du secondaire.

Bref historique

Quand Jacques Bernoulli démontre sa loi des grands nombres(2), sa motivation n'est pas purement probabiliste, mais il a déjà à l'esprit la recherche d'une méthode permettant d'*estimer* une proportion inconnue.

On considère une variable aléatoire X_n suivant une loi binomiale B (n,p). On pose $F_n = \dfrac{X_n}{n}$. Alors pour tout $\varepsilon > 0$ on a $P\left(\left|F_n - p\right| \geq \varepsilon\right) \leq \dfrac{p(1-p)}{n\varepsilon^2}$.

Et il complète ce résultat par un commentaire :

« *Je voudrais que la valeur de p que nous entreprenons de déterminer expérimentalement ne soit pas prise de façon nette et sans partage, [...], mais soit admise avec une certaine latitude , c'est-à-dire comprise entre une certaine paire de limites, pouvant être prises aussi rapprochées qu'on voudra*. »

(2) *Ars Conjectandi*, publication posthume de 1713, traduction de Robert Meunier, Irem de Rouen, 1987.

On a ici résumée toute la problématique de l'estimation par intervalle d'un paramètre inconnu.

Si on écrit de manière équivalente l'inégalité de Bernoulli :

$$P\left(F_n - \varepsilon < p < F_n + \varepsilon\right) \geq 1 - \frac{p\left(1-p\right)}{n\varepsilon^2},$$

comme $p\left(1-p\right) \leq \frac{1}{4}$, on peut en déduire :

$$P\left(F_n - \varepsilon < p < F_n + \varepsilon\right) \geq 1 - \frac{1}{4n\varepsilon^2}.$$

On obtient ainsi un encadrement aléatoire de p d'une probabilité supérieure à $1 - \frac{1}{4n\varepsilon^2}$ et d'autant plus grande que n est grand.

Par exemple, si on lance 10 000 fois une pièce de monnaie dont la probabilité inconnue de faire pile est p, le nombre de piles obtenus donne une estimation par intervalle de p d'amplitude 0,02 avec une « fiabilité » supérieure à 0,75.

Quand on réalise un grand nombre de fois cette série de 10 000 lancers, les valeurs réalisées de la variable F_n semblent fluctuer de manière désordonnée. Mais la loi de grands nombres montre que ces ***fluctuations*** sont « sous contrôle » puisqu'elles sont situées dans une certaine « fourchette » avec une probabilité dont on connaît un minorant. Il y a de l'ordre dans ce désordre apparent.

C'est la recherche d'une plus grande précision de l'estimation qui a motivé les travaux ultérieurs d'Abraham de Moivre et de Pierre-Simon de Laplace pour aboutir à l'un des théorèmes les plus importants de probabilité.

Le théorème de Moivre-Laplace

Les calculs directs sur la variable X_n sont vite compliqués quand n devient grand. Il est vrai qu'à notre époque on peut calculer avec un tableur des probabilités liées à une loi binomiale même avec des grandes valeurs de n et c'est ce qui est fait en classe de Première. Mais ces calculs ne donnent pas de résultat ayant une portée générale.

Le théorème de Moivre-Laplace apporte une information capitale sur le *comportement asymptotique* de la probabilité de fluctuation de X_n entre des bornes données.

On peut observer facilement que X_n fluctue autour de son espérance np, donc la première idée est de centrer X_n en travaillant plutôt sur $X_n - np$ dont l'espérance est nulle et donc indépendante des paramètres.

Mais si on représente sur un graphique en bâtons $X_n - np$, on s'aperçoit que quand n augmente, la dispersion de $X_n - np$ autour de 0 augmente aussi et que les bâtons s'aplatissent ce qui rend le graphique vite illisible.

En divisant $X_n - np$ par son écart type $\sqrt{np\left(1-p\right)}$, on obtient une variable

« normalisée » d'espérance 0 et de variance 1 donc plus stabilisée autour de 0.

On suppose que, pour tout entier n, la variable aléatoire X_n suit une loi binomiale B (n,p). On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$, variable centrée et réduite associée à X_n.

Alors, pour tous réels a et b tels que $a < b$, on a :

$$\lim_{n\to+\infty} P\left(a \leq Z_n \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

On peut exprimer cela en disant que la loi (discrète) de Z_n converge vers la loi normale N (0,1).

Intervalle de fluctuation asymptotique

Le théorème de Moivre-Laplace montre que si n est grand, la loi de Z_n suit approximativement une loi normale.

Cela est utilisé abondamment dans certains cursus du supérieur pour faire des calculs approchés impliquant une loi binomiale en utilisant une table de valeurs de la loi normale N (0,1).

Comme signalé précédemment, les calculs sur la loi binomiale sont faisables avec une calculatrice ou un tableur et donc cette approximation a perdu beaucoup de son intérêt. Il faut en outre remarquer que la loi binomiale est discrète et la loi normale continue, ce qui pose quelques problèmes de « correction de continuité ».

En effet, pour une variable binomiale X_n, $P\left(X_n = k\right)$ a une valeur non nulle dès lors que $0 \leq k \leq n$ alors que pour son « approximation » normale X'_n toute probabilité ponctuelle est nulle. On est donc amené à approcher $P\left(X_n = k\right)$ par la probabilité $P\left(k - 0{,}5 \leq X'_n \leq k + 0{,}5\right)$ sauf pour les valeurs extrêmes où l'on prend $P\left(X'_n \leq 0{,}5\right)$ et $P\left(n - 0{,}5 \leq X'_n\right)$.

Donc dans le secondaire cette technique n'est pas appropriée.

En revanche, on peut déduire du théorème de convergence une expression explicite d'un intervalle de fluctuation approché de X_n/n pour n grand. En première, on peut déterminer un intervalle de fluctuation exact en utilisant le tableur ou un algorithme.

Si la variable aléatoire X_n suit la loi B (n,p) alors, pour tout réel α dans]0, 1[on a :

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha,$$

où I_n désigne l'intervalle $\left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ et u_α désigne l'unique réel tel que $P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$ où Z suit la loi normale N (0,1).

L'intervalle I_n est donc un intervalle de fluctuation approché de $\frac{X_n}{n}$ au seuil $1 - \alpha$, valable pour n suffisamment grand et c'est ce qui en fait le caractère asymptotique.

Tous les exercices impliquant une loi binomiale où p est connu et où $n \geq 30$, $np \geq 5$, $n(1 - p) \geq 5$ peuvent donc se traiter avec cet intervalle de fluctuation. Si p n'est pas connu mais que l'on fait une hypothèse sur sa valeur, en particulier dans les exercices de prise de décision déjà rencontrés en seconde et en première, on peut utiliser ce nouvel intervalle de fluctuation. L'avantage en est qu'il se calcule directement sans recours au tableur. Mais on peut résoudre d'autres types d'exercice.

Une variante du problème de « surbooking » proposé dans le document ressource.

On veut construire sur un campus universitaire deux restaurants l'un de N_1 places, l'autre de N_2 places. On suppose que n étudiants prendront leur repas dans l'un de ces deux restaurants. On prévoit que les étudiants choisiront le RU1 avec la probabilité p et donc le RU2 avec la probabilité $1 - p$. Leurs choix sont indépendants. On suppose enfin que $n \geq 200$.On note X_n le nombre d'étudiants choisissant le RU1 pour un repas donné. La loi de X_n est la binomiale de paramètres n et p. À l'aide de l'intervalle de fluctuation asymptotique I_n au seuil de 95% de la variable $\frac{X_n}{n}$ on peut résoudre différents problèmes.

On peut trouver la taille minimale des deux RU pour que tous les étudiants trouvent une place dans le RU de leur choix avec une probabilité de 0,95 au moins (à 0,01 près) pour différentes valeurs de p.

Par exemple pour $p = 0{,}5$ et $n = 1000$, la taille doit être d'au moins 531 pour les deux ; pour $p = 0{,}3$ et $n = 1000$, le RU1 doit avoir au moins 329 places et le RU2 729 places. Si on fixe arbitrairement à 500 la capacité de chaque RU et que l'on suppose $p = 0{,}5$, on peut vérifier avec une inéquation du second degré que 938 étudiants pourront trouver la place qui leur convient avec une probabilité de 0,95 au moins (à 0,01 près).

Estimation d'une proportion inconnue

Bien souvent en statistique on cherche à estimer la valeur d'un paramètre inconnu lié à une certaine population. Les domaines où cette estimation est primordiale sont

extrêmement variés. Citons par exemple : la médecine au sens large (estimation de la prévalence d'une maladie, l'épidémiologie, la biologie et les recherches pharmaceutiques, …), l'écologie et l'agronomie, l'industrie (fiabilité, …), les sciences humaines, les sciences politiques, les assurances, les échanges commerciaux, … Pour une première approche de cette problématique le programme de terminale se limite à l'estimation d'une proportion car les outils probabilistes nécessaires sont disponibles.

Le premier point important est que si la population est réduite le calcul exact de la proportion peut se faire par relevé statistique direct. Mais la plupart du temps la population est trop nombreuse pour que cela soit réalisable avec un coût horaire et financier acceptable.

On a donc recours à des échantillons issus de cette population dans lesquels on fait des relevés statistiques sur le paramètre que l'on veut évaluer et on veut étendre l'information obtenue à l'ensemble de la population.

Supposons que l'on ait une population dans laquelle un certain caractère est présent dans une proportion inconnue p. Si l'on revient à la loi des grands nombres de Bernoulli, on voit que la fréquence aléatoire $\frac{X_n}{n}$ de ce caractère dans un échantillon aléatoire de taille n est un *estimateur* naturel de p d'autant plus fiable que n est grand. Que signifie précisément le terme « fiable » ?

Son espérance est p et quand un estimateur a cette propriété, on dit qu'il est sans biais. Sa variance est $\frac{p(1-p)}{n}$ et de ce fait tend vers 0 quand n tend vers l'infini ; dans ce cas on dit que l'estimateur est convergent.

Pour des raisons pratiques évidentes, on ne peut pas multiplier les échantillons visant à estimer un même paramètre. Pour un essai thérapeutique par exemple, il est souvent très difficile et trop coûteux en temps de refaire plusieurs études.

Le problème est donc de pouvoir donner avec un seul échantillon une estimation du paramètre avec une quantification précise du *risque* que l'on prend à faire *confiance* à cet échantillon (les statisticiens parlent parfois de *pari* …).

Mais le monde dans sa complication infinie est décidemment bien fait[3] car on l'a déjà vu : les fluctuations de $\frac{X_n}{n}$ bien qu'apparemment erratiques sont bornées de manière très précise grâce à l'intervalle de fluctuation asymptotique. Les bornes de cet intervalle dépendent de n, de u_α (lié à la précision) mais aussi malheureusement du paramètre inconnu p. La manipulation de la double inégalité faite au début à partir de la loi des grands nombres visant à encadrer p par deux variables aléatoires n'apporte donc ici aucune information utilisable directement.

(3) Voir à ce sujet dans le document ressource la conviction religieuse que Moivre (mais pas Laplace) en tire.

Mais en acceptant de perdre un peu de précision on peut obtenir un intervalle aléatoire qui contient p avec une probabilité pouvant être minorée. Comme $p(1-p) \leq \frac{1}{4}$ pour toute proportion p, on peut élargir l'intervalle en un intervalle plus simple. Si on décide de fixer le seuil de confiance à 0,95, on a alors $u_\alpha \approx 1{,}96$ (grâce aux propriétés de la loi normale) que l'on peut lui aussi majorer par 2 (on n'est plus à cela près…) et on obtient l'intervalle de confiance très simple $\left(F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}\right)$.

A priori cet intervalle a une probabilité voisine de 0,95 de contenir p si n est grand. Mais on préfèrerait avoir un vrai niveau de confiance c'est-à-dire avoir l'assurance que l'intervalle a une probabilité minimale de contenir p.

On peut démontrer (voir le document ressource) qu'il existe un entier n_0 tel que si $n \geq n_0$, $\left(F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}\right)$ a une probabilité d'au moins 0,95 de contenir p.

Cet entier n_0 dépend de p et est d'autant plus grand que p est proche de 0,5 car c'est pour cette valeur que la variance de $F_n = \frac{X_n}{n}$ est maximale (n_0 varie de 31 pour $p = 0{,}35$ à 529 pour $p = 0{,}5$)[4].

On dispose maintenant d'un intervalle *aléatoire* contenant le paramètre inconnu p avec une probabilité d'au moins 0,95 si n est assez grand.

En pratique, on extrait un échantillon au hasard et théoriquement avec remise[5] (c'est uniquement à ce stade qu'il y a de l'aléatoire). Il est important que tous les éléments de la population aient la même probabilité d'être présents dans l'échantillon pour pouvoir appliquer le modèle binomial à la base de cette théorie[6].

On relève la fréquence du caractère étudié dans l'échantillon qui constitue ainsi une réalisation de la variable $F_n = \frac{X_n}{n}$ et on en déduit un intervalle de confiance de p réalisé au niveau de confiance 0,95.

Si par exemple on obtient l'intervalle de confiance réalisé [0,45 ; 0,48] on ne peut bien entendu pas dire que p appartient à cet intervalle avec une probabilité d'au moins 0,95 car cette appartenance est vraie ou pas puisque p bien qu'inconnu est fixé : il n'y a plus d'aléatoire à ce stade. Simplement on peut accorder un niveau de confiance de 0,95 à l'affirmation « p appartient à [0,45 ; 0,48] », le risque de se tromper est lui limité à 0,05. Cette confiance et ce risque sont des « réminiscences »

(4) La minoration par 0,95 pour $n \geq 30$ et $0{,}2 \leq p \leq 0{,}8$ indiquée dans le programme de seconde n'est donc pas tout à fait exacte.

(5) En pratique c'est plutôt sans remise et la loi est hypergéométrique mais si la population est grande il n'y a guère de différence...

(6) Voir le document ressource pour plus de précisions sur les techniques d'échantillonnage.

de la probabilité liée à la procédure : si on réalisait 100 fois l'échantillonnage on obtiendrait 100 intervalles de confiance réalisés différents et en principe environ 95% d'entre eux contiennent p. 95% n'est pas une probabilité mais plutôt une valeur statistique.
On a obtenu un intervalle de confiance de la proportion inconnue p au niveau 0,95 très simple valable pour toutes les situations dès lors que n est suffisamment grand (529 si on veut être rigoureux).

On peut se demander si on n'aurait pas une meilleure précision en ne procédant pas aux deux élargissements de l'intervalle.

On peut démontrer(7) que la probabilité de l'événement

$$\left[\mathrm{F}_n - \frac{u_\alpha}{\sqrt{n}}\sqrt{\mathrm{F}_n\left(1-\mathrm{F}_n\right)} \leq p \leq \mathrm{F}_n + \frac{u_\alpha}{\sqrt{n}}\sqrt{\mathrm{F}_n\left(1-\mathrm{F}_n\right)}\right]$$

tend vers $1-\alpha$ quand n tend vers l'infini. On obtient ainsi un intervalle de confiance de p au niveau $1-\alpha$.

Il est certainement plus précis que le précédent mais ce gain de précision n'est pas vraiment spectaculaire.

Si sur un échantillon de taille 1000 on observe une valeur de la fréquence égale à 0,48, l'intervalle de confiance réalisé vu plus haut est [0,448 ; 0,512] et celui obtenu avec cette nouvelle formule est [0,449 ; 0,511].

Pour une fréquence observée de 0,15, on aurait [0,118 ; 0,182] pour le premier et [0,128 ; 0,172] pour le deuxième.

Approche graphique de la notion d'intervalle de confiance

On peut visualiser sous GeoGebra la construction d'un intervalle de confiance à partir de l'intervalle de fluctuation asymptotique. Pour cela, on trace les courbes des deux fonctions de p :

$$f_1 : p \mapsto p - u_\alpha \frac{\sqrt{p\left(1-p\right)}}{\sqrt{n}} \text{ et } f_2 : p \mapsto p + u_\alpha \frac{\sqrt{p\left(1-p\right)}}{\sqrt{n}}$$

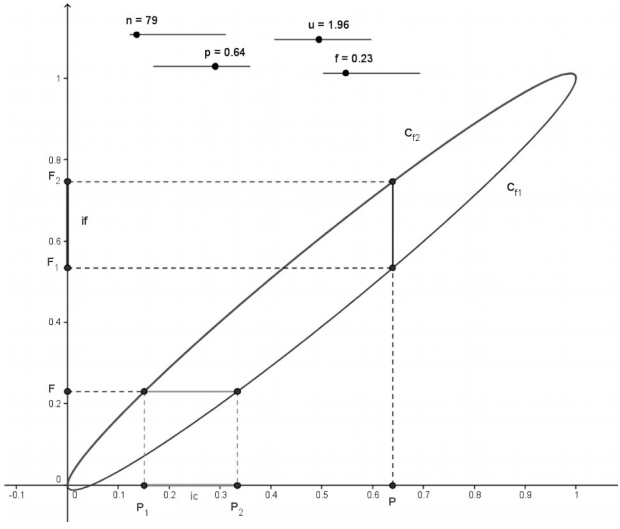
pour une valeur de n et une valeur de u_α fixées (mais modifiables par des curseurs).

Pour une valeur de p donnée, on obtient en ordonnée deux valeurs $f_1(p)$ et $f_2(p)$ qui représentent les bornes de l'intervalle de fluctuation pour cette valeur de p.

Inversement, en se donnant une valeur f, on peut faire tracer en abscisses l'intervalle $\left[p_1, p_2\right]$ dont les bornes sont les antécédents de f par les deux fonctions. L'intervalle $\left[p_1, p_2\right]$ est l'ensemble des valeurs de p solutions de l'inéquation $f_1(p) \leq f_2(p)$ et celui-ci constitue une réalisation de l'intervalle de confiance aléatoire $\left[f_2^{-1}\left(\mathrm{F}_n\right), f_1^{-1}\left(\mathrm{F}_n\right)\right]$ (dont la probabilité tend vers $1-\alpha$).

(7) Grâce au théorème de Slutsky.

Pour $n = 1000$, $f = 0{,}15$, on trouve $\left[p_1, p_2\right] \approx [0{,}129\ ;\ 0{,}173]$ à comparer avec [0,128 ; 0,172] obtenu précédemment.



On peut calculer explicitement $f_1^{-1}\left(\mathrm{F}_n\right)$ et $f_2^{-1}\left(\mathrm{F}_n\right)$. Les expressions obtenues sont un peu compliquées mais on obtient $\mathrm{F}_n + \dfrac{u_\alpha}{\sqrt{n}}\sqrt{\mathrm{F}_n\left(1-\mathrm{F}_n\right)}$ et $\mathrm{F}_n - \dfrac{u_\alpha}{\sqrt{n}}\sqrt{\mathrm{F}_n\left(1-\mathrm{F}_n\right)}$ comme valeurs approchées en négligeant les termes en $\dfrac{1}{n}$, ce qui rejoint l'intervalle donné auparavant.

Pour aller plus loin sur l'estimation

La seule estimation au programme en terminale est celle d'une proportion mais il y a bien sûr beaucoup d'autres paramètres que l'on cherche à estimer : moyennes, écart-type, nombre d'individus d'une population...

Quelques exemples de ces estimations seront proposés en ligne sur le site de l'APMEP.

Bibliographie

Ouvrage fondateur de lecture très accessible : Pierre-Simon de Laplace, Œuvres complètes disponibles à l'adresse http://gallica.bnf.fr/

Ouvrage théorique : Mathématiques appliquées L3 Pearson éducation

Ouvrage pratique de statistiques appliquées à la médecine : Jean Bouyer Inserm de boeck estem

Ouvrage pratique de statistiques appliquées à la psychologie : Béatrice Beaufils Tome 2 Lexifac