METHODES EN STATISTIQUE : ESTIMATION

HENRY Michel
IREM de BESANÇON

I - LE PROBLEME DE L'ESTIMATION

A) LES ORIGINES ; CONCEPTIONS DE JACQUES BERNOULLI.

B) L'ESPRIT FREQUENTISTE DES PROGRAMMES DU SECOND DEGRE

C) PROBLEME D'AJUSTEMENT D'UNE LOI

II - MODELE DE LA STATISTIQUE INFERENTIELLE

A) MODELE PROBABILISTE

B) MODELE DE LA STATISTIQUE

C) MODELE POUR L'ECHANTILLONNAGE

III - ECHANTILLONNAGE ET STATISTIQUES

A) LOI D'UN ECHANTILLON ET FONCTION DE VRAISEMBLANCE

B) RESUME STATISTIQUE ET LOI D'UNE STATISTIQUE

C) STATISTIQUES USUELLES $\overline{X}$ ET S^2
CAS PARTICULIER D'UN ECHANTILLON GAUSSIEN

IV - ESTIMATEURS

A) CADRE DE L'ESTIMATION

B) QUALITES D'UN ESTIMATEUR : **biais, convergence, efficacité, limite théorique des performances d'un estimateur**

C) RECHERCHE D'UN ESTIMATEUR : METHODE DU MAXIMUM DE VRAISEMBLANCE

V - ESTIMATION PAR INTERVALLE

A) POSITION DU PROBLEME

B) PRINCIPE DE LA DETERMINATION D'UN INTERVALLE DE CONFIANCE

C) MISE EN PRATIQUE DE LA RECHERCHE DE L'INTERVALLE DE CONFIANCE

D) FORMULATION DE LA CONCLUSION

I - LE PROBLEME DE L'ESTIMATION

A) LES ORIGINES ; CONCEPTIONS DE JACQUES BERNOULLI.

Dans une approche naïve du Calcul des Probabilités, on cherche à évaluer les "chances" que l'on a d'observer un événement dont l'apparition est conditionnée par le hasard.

Dans une démarche scientifique, il s'agit de discerner les variables pertinentes dans des situations concrètes permettant de dégager un modèle suffisamment simple, dans lequel il est possible d'utiliser les outils mathématiques.

Cette intention de calculer des probabilités suppose une prise de parti épistémologique qui n'a été ni simple ni précoce dans l'histoire des sciences. Il a fallu accepter l'idée que les phénomènes aléatoires sont des phénomènes objectifs qui peuvent être quantifiés indépendamment de la sensibilité et des préjugés de l'observateur (position objectiviste).

Dans ce cadre, le calcul des probabilités s'accommode-t-il du déterminisme ?

Dans la préface à l'"*Essai philosophique sur les probabilités*" de Laplace, René Thom indique [1, p.23] : "*En science, l'aléatoire pur, c'est le processus markovien, où toute trace du passé s'abolit dans la genèse du nouveau coup... l'aléatoire pur exige un fait sans cause, c'est-à-dire un commencement absolu.*"

Ainsi, dans la dyade déterminisme - hasard, une position philosophique interprétant les phénomènes naturels comme purement aléatoires, ou soumis à la volonté divine imprévisible, conduit nécessairement l'observateur à exclure tout modèle mathématique et le réduit à l'attribution de "probabilités subjectives" (selon les termes repris par Poincaré [2, p.195]) qui ne peuvent, en l'absence de préférences, qu'être réduites à l'équiprobabilité pour les différentes éventualités qui sont susceptibles de se présenter : l'âne de Buridan aurait une chance sur deux de choisir d'abord le seau d'eau - s'il ne veut pas mourir de faim et de soif.

Dans son livre "*Au hasard*", Ivar Ekeland nous livre de belles pages sur l'approche moderne du hasard, évoquant l'évolution historique des conceptions qui conduisent au traitement scientifique [4, p.63] : "*Si la réalité ultime est décrite par le calcul des probabilités, le monde sera soumis aux lois de la statistique.*" Et, [p.80] : "*l'incertitude est une des données fondamentales de l'histoire humaine et de notre vie quotidienne. En permanence, il nous faut prendre des décisions dans un contexte que nous apprécions mal*".

Cette nécessité de prendre des décisions avait conduit Pascal, dans sa correspondance avec Fermat, à retenir comme concept de probabilité ce qui est a priori calculable par logique, symétries et dénombrements, dans le degré d'incertitude que l'on a sur l'apparition de tel ou tel résultat d'une expérience aléatoire clairement décrite et dont les événements élémentaires sont des cas équiprobables (problème des Partis : [6, p.217]). D'où la formule donnant la probabilité d'un événement composé :

$$\frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}$$

qui semble exclure toute subjectivité, mais qui ne s'applique qu'aux expériences aléatoires bien caractéristiques (dés, urnes...) ayant un petit nombre d'issues que l'on peut, selon les termes de Laplace [1, p.35] : "*réduire à un certain nombre de cas également possibles*".

La modélisation (par exemple en termes d'urnes) de telles expériences est alors élémentaire et le calcul des probabilités est essentiellement du ressort mathématique. D'où le terme de "géométrie du hasard" introduit par Pascal, qu'il faut comprendre dans le sens où la géométrie est la "mathématique du réel" qui, du point de vue de Platon, s'opposerait à l'arithmétique opérant dans le domaine des Idées.

Mais cette conception est insuffisante pour s'attaquer aux vrais problèmes et est marquée d'un vice épistémologique à la base, comme le fait remarquer Poincaré [2, p.192] : "*On est donc réduit à compléter cette définition [de Pascal] en disant* : «... au nombre total des cas possibles, pourvu que ces cas soient également probables». *Nous voilà donc réduits à définir le probable par le probable.*"

Comment, en réalité, relier le "degré d'incertitude" (terme de Laplace) que nous avons sur un événement à la probabilité objective de cet événement, que seule la compilation statistique dans un grand nombre de réalisations peut révéler a posteriori (Poincaré) ?

Cette question, qu'en termes actuels nous appelons "problème d'estimation des probabilités", était déjà posée en toute clarté par Jacques Bernoulli dans son *Ars Conjectandi*, publié en 1713 après sa mort ([5, p.40]). Il souligne les conditions restrictives (jeux de hasard,...) auxquelles s'applique la définition de Pascal et montre que la complexité des phénomènes naturels suppose une autre manière de concevoir la probabilité qui, pour ne pas être subjective, doit provenir d'une étude des fréquences des événements issus de nombreuses expériences ou situations aléatoires identiques. Cette assimilation fréquence - probabilité est justifiée par la loi des grands nombres qui, d'une perception intuitive commune, passe au statut de résultat mathématiquement établi : c'est l'objet de l'*Ars Conjectandi*.

Il est saisissant de voir en quels termes contemporains Bernoulli introduit cette problématique [5, p.42,44] :

« On en est ainsi venu à ce point que pour former selon les règles des conjonctures sur n'importe quelle chose il est seulement requis d'une part que les nombres de cas soient soigneusement déterminés, et d'autre part que soit défini combien les uns peuvent arriver plus facilement que les autres. Mais c'est ici enfin que surgit une difficulté, nous semble-t-il : cela peut se voir à peine dans quelques très rares cas et ne se produit presque pas en dehors des jeux de hasard que leurs premiers inventeurs ont pris soin d'organiser en vue de se ménager l'équité, de telle sorte que fussent assurés et connus les nombres de cas qui doivent entraîner le gain ou la perte, et de telle sorte que tous ces cas puissent arriver avec une égale facilité. En effet lorsqu'il s'agit de tous les autres résultats, dépendant pour la plupart soit de l'oeuvre de nature soit de l'arbitre des hommes, cela n'a pas du tout lieu. Ainsi, par exemple, les nombres de cas sont connus lorsqu'il s'agit des dés, car pour chacun des dés les cas sont manifestement aussi nombreux que les bases, et ils sont tous également enclins à échoir. ...
Mais qui donc parmi les mortels définira par exemple le nombre de maladies,... ?
Qui encore recensera les cas innombrables des changement auquel l'air est soumis chaque jour, en sorte qu'on puisse à partir de là conjecturer ce que sera son état après un mois ?...
En outre qui aurait sur la nature de l'esprit humain, ou sur l'admirable fabrique de notre corps une vue suffisante pour oser déterminer dans les jeux, qui dépendent en totalité ou en partie de la finesse de celui-là ou de l'agilité de celui-ci, les cas qui peuvent donner la victoire ou l'échec. ...

Mais à la vérité ici s'offre à nous un autre chemin pour obtenir ce que nous cherchons. Ce qu'il n'est pas donné d'obtenir a priori l'est du moins *a posteriori*, c'est-à-dire qu'il sera possible de l'extraire en observant l'issue de nombreux exemples semblables ; car on doit présumer que, par la suite, chaque fait peut arriver et ne pas arriver dans le même nombre de cas qu'il avait été constaté auparavant, dans un état de choses semblables, qu'il arrivait ou n'arrivait pas. »

B) L'ESPRIT FREQUENTISTE DES PROGRAMMES DU SECOND DEGRE

(cf. les programmes de 1ère et de terminales de 1991 et l'article de REPERES-IREM [9])

A partir de ces considérations épistémologiques, nous pouvons analyser la démarche de ces programmes qui s'affirment plus "fréquentiste" que les précédents dans l'approche de la notion de probabilité.

Modélisation d'une expérience aléatoire

Ces programmes inscrivent d'abord l'introduction des probabilités dans la continuité des apprentissages en statistique, particulièrement lors de l'étude des séries statistiques réalisée en seconde.

Mais, et ceci est fondamental, une probabilité ne peut avoir de sens que si elle concerne un événement associé à une expérience aléatoire.

En l'absence d'expérience aléatoire, pas de probabilité !

Le concept de probabilité suppose donc acquis ceux d'expérience aléatoire et d'événement. C'est d'ailleurs souligné en premier lieu dans le programme : "*l'objectif est d'entraîner les élèves à décrire quelques expériences aléatoires simples*". Ce concept se dégagera alors de cette activité de description portant sur différents exemples dans des cadres variés.

Si, ensuite, on veut développer un travail mathématique, la notion d'expérience aléatoire doit pouvoir être modélisée. Nous savons que le langage et les opérations sur les ensembles le permettent effectivement. Nous ferons les trois hypothèses suivantes :

a- Une telle expérience met en jeu un phénomène aléatoire : son issue ne peut être prévue à l'avance, elle est le "fruit du hasard". Ainsi, les conditions de l'expérience, telles qu'elles sont décrites, ne déterminent pas l'un des résultats possibles de manière absolue.

b- Pour représenter une expérience aléatoire, on considère donc un ensemble de résultats possibles, bien identifiés.

c- Enfin, une expérience aléatoire doit être reproductible dans les mêmes conditions (au moins par la pensée).

Ainsi, les faits historiques ne peuvent être considérés comme résultant d'expériences aléatoires et ne sauraient être probabilisés.

A la notion d'expérience aléatoire est donc liée celle de hasard, avec les difficultés épistémologiques ou philosophiques que nous connaissons et qui justifient une longue dissertation en introduction de l'ouvrage de Henri Poincaré "*Calcul des probabilités*" publié en 1912. En voici un extrait :

Le déterminisme Laplacien pose que "*le mot hasard est tout simplement un synonyme d'ignorance, qu'est-ce que cela veut dire ? ... Il faut donc bien que le hasard soit autre chose que le nom que nous donnons à notre ignorance, que parmi les phénomènes dont nous ignorons les causes, nous devions distinguer les phénomènes fortuits, sur lesquels le calcul des probabilités nous renseignera provisoirement, et ceux qui ne sont pas fortuits et sur lesquels nous ne pouvons rien dire, tant que nous n'aurons pas déterminé les lois qui les régissent*" [3, p.3].

Cette remarque montre que pour comprendre les difficultés conceptuelles de la notion de probabilité, nous nous heurterons à des obstacles épistémologiques et l'on n'évitera pas un travail de fond avec les élèves qui abordent l'aléatoire pour la première fois en mathématiques

Ces obstacles nous sont révélés par les hésitations historiques des Bernoulli, D'Alembert, Laplace, Poincaré. Il est bon, alors, que les enseignants de mathématiques aient fait le point sur leurs propres conceptions, les aient confrontées à celles des mathématiciens du passé pour ne pas être pris au dépourvu par tel ou tel comportement d'élève qui rencontre à cette étape des difficultés de compréhension.

Approche fréquentiste de la probabilité

C'est ici un objectif particulier du programme que d'associer la notion de probabilité d'un événement à la fréquence stabilisée de cet événement au cours d'un grand nombre d'expériences identiques. Le programme de première ajoute : "*Pour introduire la notion de probabilité, on s'appuiera sur l'étude des séries statistiques obtenues par répétition d'une expérience aléatoire, en soulignant les propriétés des fréquences et la relative stabilité de la fréquence d'un événement donné lorsque cette expérience est répétée un grand nombre de fois.*"

Les spécialistes y verront une allusion à la loi des grands nombres, tout en pesant les difficultés épistémologiques qui lui sont inhérentes et les difficultés didactiques d'un énoncé précis à ce niveau.

La probabilité est ainsi liée à la notion de fréquence, laquelle prend du sens par la description de multiples exemples de phénomènes aléatoires. C'est pour cela que nous parlons de conception fréquentiste, en opposition à la conception qui postule l'équiprobabilité des événements élémentaires, posée a priori pour des raisons de symétrie et que nous appellerons "*l'approche pascalienne*", en référence à l'auteur de la formule de définition de la probabilité dans la "*géométrie du hasard*".

La probabilité est donc conçue dans les programmes du secondaire comme une valeur numérique injectée dans un modèle de l'expérience sensible, modèle dégagé de l'activité qui, en statistique, est centrée sur le recueil et l'organisation de données.

Remarquons que cette démarche est cohérente avec les objectifs des programmes de collège et de seconde en vigueur, privilégiant l'activité des élèves, l'observation et la formulation de conjectures, avant que le modèle mathématique soit institutionnalisé. S'opposant à une présentation formelle des mathématiques, ces programmes mettent en avant l'étude d'outils de description d'une réalité concrète.

Ces questions ayant été clarifiées, on est placé devant un problème d'estimation : dans quelle mesure les 48,5 % de "pile" observés justifient-ils le 0,5 attribué à la probabilité du "pile" ? Problème pas très difficile au niveau BTS, mais exclu en classe de première sur le plan conceptuel même. Au fait, pourquoi 0,5 et non 0,49 ? (la pièce est peut-être déséquilibrée).

Ainsi la fréquence observée sur un grand nombre d'expériences ne doit pas être confondue avec la notion mathématique de probabilité. Celle-ci est conçue comme donnée numérique du modèle, estimée à partir de l'observation de la stabilité de cette fréquence.

Dans sa cohérence, le programme propose "*l'observation de la stabilité approximative de la fréquence f_n d'un événement donné lorsque l'expérience est répétée un grand nombre n de fois*".

Il ne s'agit donc pas d'établir une estimation de la probabilité p limite. D'ailleurs, dans ce cas, la meilleure estimation est la fréquence observée pour le plus grand nombre d'expériences et l'on n'a que faire de sa stabilisation.

De plus, et le programme le précise, cette stabilisation ne peut être que relative. En effet, en théorie, la convergence de f_n vers p n'est pas monotone :
si $p = ½$, il y a une chance sur 2 pour qu'au prochain tirage f_n s'éloigne de p.

Il convient cependant d'habituer les élèves à cette stabilité relative, de leur faire apprécier sur des exemples le nombre n d'expériences nécessaires pour l'observer et de leur donner confiance dans l'approche fréquentiste. Celle-ci, en effet, correspond mieux à la pratique sociale et à l'usage actuel en statistique dans l'induction suivante : si, dans un échantillon assez vaste pris au hasard dans une population, j'observe une proportion p d'éléments de catégorie A, le choix au hasard d'un autre élément de la population donnera A avec une probabilité que je peux prendre égale à p.

Cette pratique imagine la répétition de l'expérience une fois de plus et suppose la stabilité de la fréquence observée dès que l'échantillon préalable est assez grand ([1]).

Le programme propose donc de telles observations, concrètement, en poursuivant l'objectif de privilégier l'activité des élèves comme support à leur conceptualisation.

c) PROBLEME D'AJUSTEMENT D'UNE LOI

La démarche empirique qui consiste à "estimer" une probabilité par la fréquence observée de l'événement reste très limitée dans ses résultats pratiques.

D'une part, elle ne dit pas quel est le degré d'approximation qu'on obtient ainsi, ni avec quelle précision on peut introduire les probabilités dans des calculs complexes, où des incertitudes sur les données initiales peuvent provoquer de grandes variations sur les résultats.

D'autre part, elle limite le concept de probabilité ou son domaine d'application aux expériences aléatoires effectivement répétables un grand nombre de fois et exclut de son champ les situations complexes non reproductibles, économiques par exemple.

---

1) voir cependant Emile BOREL : *Valeur pratique et philosophie des probabilités*, Gauthier-Villars, Paris, 2ème éd. 1952.

Le développement de la théorie des probabilités a montré l'importance de la notion de variable aléatoire et, pour le développement des modèles, de celle de loi.

Ainsi, plutôt que de déterminer a posteriori toutes les probabilités relatives aux événements issus d'une expérience aléatoire, il est plus avantageux de décrire cette expérience en introduisant les variables pertinentes et de déterminer les lois de ces variables qui permettent ensuite tous les calculs de probabilités souhaités.

Dans la pratique, la loi est un outil théorique sensé décrire au mieux la répartition des valeurs possibles de la variable aléatoire. Le problème est donc de choisir, parmi un ensemble catalogué de lois théoriques bien connues, celle qui satisfait le mieux aux conditions de l'expérience aléatoire. C'est ce qu'on appelle un problème d'ajustement.

En réalité, suivant l'expérience aléatoire et le choix des variables la décrivant, il y a certains types de lois qui viennent s'imposer. Les connaissances et le savoir-faire du probabiliste - statisticien lui permettront de faire le bon choix.

Par exemple, s'il pleut et si, sur un territoire limité, je m'intéresse à la quantité d'eau reçue par les éléments de surface du sol, il sera naturel de suggérer une loi uniforme en dimension 2.

Par contre si j'arrose mon gazon avec un jet fixe muni d'une pomme de dispersion, je penserai plutôt à une loi normale que j'essaierai de contrôler expérimentalement .

Ainsi, l'outil mathématique puissant que donne le modèle de Kolmogorov, dans lequel la notion de loi est centrale, nous conduit-il à estimer non plus la probabilité de chaque événement, mais les paramètres qui déterminent numériquement les lois des variables en cause.

II - MODELES DE LA STATISTIQUE INFERENTIELLE

(Pour cet enseignement à des sections de techniciens supérieurs, on pourra se reporter à [8] et [10])

A) MODELE PROBABILISTE

En statistique inférentielle, on s'intéresse à des expériences aléatoires particulières et à un problème particulier :
Etant donnée une population $\mathcal{P}$ que l'on désire étudier, l'expérience consiste en un prélèvement (non exhaustif, i.e. avec remise, en théorie) "au hasard" de n éléments de $\mathcal{P}$ (l'échantillon au sens commun).

Le problème est d'inférer, à partir de l'observation de cet échantillon, les propriétés de $\mathcal{P}$.

On va supposer pour la suite que ces propriétés sont interprétées par les valeurs d'un caractère χ (éventuellement multidimensionnel) que nous prendrons quantitatif (pour nous limiter).

Le modèle probabiliste décrivant le prélèvement au hasard d'un élément de $\mathcal{P}$ introduit un ensemble Ω (avec les notations habituelles) représentant cette population :

$$(\Omega, \mathcal{T}, P) \xrightarrow{X} (\mathbf{R}^d, \mathcal{B}, P_X)$$

avec la signification des lettres :
- Ω ensemble des éléments de la population, • $\mathcal{T}$ tribu sur Ω, • P distribution de la probabilité sur Ω
- X v.a. représentant le caractère χ, à valeurs dans $\mathbf{R}^d$, • $\mathcal{B}$ boréliens de $\mathbf{R}^d$,
- P_X mesure image de P par X décrite généralement par l'un des moyens suivants :
  - probabilités élémentaires discrètes
  - fonction de répartition
  - densité
  - fonction caractéristique

B) MODELE DE LA STATISTIQUE

Le modèle statistique que nous allons utiliser doit décrire le prélèvement de l'échantillon.

Plaçons nous directement dans l'espace image de $\mathcal{P}$ par le caractère χ.

Soit E_0 l'ensemble des valeurs possibles de χ. On introduit X_0 la v.a.(2) représentant le tirage au hasard d'un élément de $\mathcal{P}$ et l'application à cet élément de χ (observation de la valeur du caractère). X_0 dans ce modèle est alors l'identité de E_0, mais son introduction assure la prise en compte du caractère aléatoire des valeurs observées.

Les valeurs de X_0 sont réparties suivant une loi de probabilité P_{X_0}. X_0 sera appelée variable parente (de l'échantillon) et P_{X_0} la loi parente.

En réalité, c'est P_{X_0} que l'on désire connaître. Il y a divers degrés d'incertitude sur cette loi qui déterminent la répartition des valeurs de χ et par conséquent les éléments caractéristiques de $\mathcal{P}$ comme les valeurs moyenne, écart-type, etc...

Suivant la connaissance de $\mathcal{P}$ et de χ que l'on a, P_{X_0} sera à rechercher parmi une famille plus ou moins grande de lois possibles, chacune déterminée numériquement par différents paramètres.

On indique cette indétermination par l'introduction d'un paramètre $\theta \in \Theta$ qui peut être numérique, éventuellement multidimensionnel. Il caractérise donc le couple $(\mathcal{P}, \chi)$.

P_{X_0} dépend de θ, on la note alors $P_{X_0,\theta}$ et l'ensemble des lois possibles par $(P_{X_0,\theta})$, $\theta \in \Theta$.

D'où le modèle représentant le problème de l'inférence : $\left(E_0, \mathcal{B}_0, (P_{X_0,\theta})\ \theta \in \Theta\right)$.

C) MODELE POUR L'ECHANTILLONNAGE

L'expérience consiste à observer (et traiter) n réalisations indépendantes de X_0, prélèvements au hasard avec remise, c'est-à-dire n observations de χ.

On notera X_i la i-ème opération de ce type :

X_i est une v.a. définie sur E_0 (identité) de même loi que X_0.

La réplique n fois de la même expérience aléatoire dans les mêmes conditions est interprétée dans le modèle par l'hypothèse que les X_i sont indépendantes.

Soit alors $E = E_0^n$ l'ensemble des n-uples d'éléments de E_0 et $X = (X_1,...,X_n)$.

X est une v.a. définie sur E ; on l'appelle le n-échantillon de X_0. X est caractérisé par le fait que les X_i sont de même loi (que X_0) et indépendantes.

La loi de X est entièrement déterminée par $P_{X_0,\theta}$ et par l'hypothèse d'indépendance. On la désigne par $P_{X,\theta}$. Pour les spécialistes, c'est le produit tensoriel de n mesures identiques à celle qui représente $P_{X_0,\theta}$. Notamment, si $P_{X_0,\theta}$ est représentée par une densité $f_{X_0}(x_0,\theta)$, alors X aura pour densité :

$$f_X(x,\theta) = \prod_{i=1}^{n} f_{X_0}(x_i,\theta)$$

Le modèle de l'échantillonnage est ainsi : $(E, \mathcal{B}, (P_{X,\theta})\ \theta \in \Theta)$.

L'estimation consiste donc à déterminer θ à partir d'une observation de X, qui est un n-uple de valeurs $x=(x_1,...,x_n) \in E$ obtenu par prélèvement de n éléments de $\mathcal{P}$. On le note avec des lettres minuscules.

La théorie de l'estimation peut être insérée dans un modèle plus vaste : le modèle de la décision qui s'applique aussi bien en théorie des tests d'hypothèses ou en théorie des jeux. Sa présentation est un peu lourde et n'est pas nécessaire ici.

---

2) v.a. : abréviation de variable aléatoire.

III - ECHANTILLONNAGE ET STATISTIQUE

A) LOI D'UN ECHANTILLON ET FONCTION DE VRAISEMBLANCE

Dans le cadre de la théorie de la mesure, $P_{X,\theta}$ est considérée comme une mesure sur E. Dans la pratique courante, il y a deux situations bien différentes :
- X est discrète et $P_{X,\theta}$ est alors donnée par la liste des probabilités élémentaires,
- X est continue et, le plus souvent, $P_{X,\theta}$ est donnée par sa densité par rapport à la mesure de Lebesgue (uniforme) sur E, muni de la tribu des boréliens.

Donnons des exemples :

* n-échantillon de Bernoulli

$$X_0 = \begin{cases} 1 & \text{avec la probabilité } p \\ 0 & \text{avec la proba } q = 1-p \end{cases}, \quad p \text{ inconnu}$$

alors $P_{X_0,p}$ est déterminée par les valeurs $p^{x_0}.q^{1-x_0}$ des probabilités associées aux valeurs x_0 de X_0.
$X = (X_1,...,X_n)$ de loi $P_{X,p}$ donne $f_X(x,p) = p^{\Sigma x_i}.q^{n-\Sigma x_i}$ pour la probabilité associée à la valeur $x = (x_1,...,x_n)$ de l'échantillon.

* n-échantillon de Poisson :

X_0 à valeurs dans **N** avec $P(X_0 = x_0) = \dfrac{\lambda^{x_0}.e^{-\lambda}}{x_0!}$, λ inconnu

La loi de X est donnée par les probabilités $f_X(x,\lambda) = \dfrac{\lambda^{\Sigma x_i}.e^{-n\lambda}}{\prod_{i=1}^{n}(x_i!)}$

* n-échantillon normal :

La v.a. X_0 de loi $\mathcal{N}(m, \sigma^2)$ a pour densité $f_{X_0}(x_0,\theta) = \dfrac{1}{\sqrt{2\pi}.\sigma}.e^{-\frac{1}{2}\left(\frac{x_0-m}{\sigma}\right)^2}$, $\theta = (m,\sigma)$ inconnu

La loi de l'échantillon X : $P_{X,\theta}$ est donnée par la densité : $f_X(x,\theta) = \dfrac{1}{(2\pi)^{n/2}\sigma^n}.e^{-\frac{1}{2}\Sigma\left(\frac{x_i-m}{\sigma}\right)^2}$

par rapport à la mesure de Lebesgue sur $\mathbf{R}^n$.

Aussi bien dans le cas discret que continu, les fonctions $f_X(x,\theta)$ associées à la loi $P_{X,\theta}$ sont appelées fonctions de vraisemblance du modèle.

B) RESUME STATISTIQUE ET LOI D'UNE STATISTIQUE

Dans la pratique, n peut être grand et la succession des valeurs observées de l'échantillon est lourde et peu lisible. Comme en statistique descriptive, on résume l'ensemble des valeurs observées par des paramètres caractérisant leur position et leur dispersion (statistiques d'ordre, de position : médiane, moyenne ; de dispersion : variance, écart-type ou autres résumés plus compliqués).

Cela revient à résumer l'information contenue dans X par un ou plusieurs nombres significatifs, calculés à partir des valeurs observées.

Pour décrire cela, on complète le modèle de l'échantillonnage par l'introduction d'une application $T : E \to \mathbf{R}$ (ou $\mathbf{R}^2$,...) qui, composée avec X, est alors une variable aléatoire :

$$E \xrightarrow[\text{aléatoire}]{X} E \xrightarrow[\text{statistique}]{T} \mathbf{R}$$ dont la loi sur **R** est $P_{T,\theta}$.

Les valeurs observées t de T sont en effet le fruit du tirage au hasard de l'échantillon.

D'où le modèle : $(\mathcal{P}^n, \mathcal{I}, P_\theta) \xrightarrow{X} (E, \mathcal{B}, P_{X,\theta}) \xrightarrow{T} (\mathbf{R}, \mathcal{B}_R, P_{T,\theta})$

modèle abstrait tirage de l'échantillon	modèle de la statistique	modèle de l'inférence

Ainsi l'observation t de T donnera des renseignements sur θ (i.e. la loi de X_0), à condition de connaître $P_{T,\theta}$ à partir de $P_{X,\theta}$ et T (mesure image). C'est tout le problème de la statistique inférentielle : savoir calculer $P_{T,\theta}$.

T est appelée en général une statistique. Dans le cadre de l'estimation [$E = E_o^n$ et $X = (X_1,...,X_n)$], on dit aussi que T est un estimateur de θ.

c) Statistiques usuelles $\bar{X}$ et S^2

Dans l'étude des séries statistiques issues d'un caractère quantitatif, on a vu le rôle important joué par la moyenne $m = {}^1/_N \Sigma x_i$ de la population comme paramètre de position, et par la variance $\sigma^2 = {}^1/_N \Sigma(x_i - m)^2$ comme paramètre de dispersion (où N est la taille de la population).

En probabilités, on constate que lorsque l'on connaît le type de loi d'une v.a., dans les cas les plus fréquents, le paramètre qui détermine complètement cette loi est connu dès que l'on connaît son espérance mathématique (Bernoulli, binomiale, Poisson, Pascal, hypergéométrique, exponentielle, Γ) ou en outre son écart-type (uniforme, normale).

Pour connaître la loi parente de X_0, si elle est de ces types, il suffit donc de déterminer des valeurs assez précises pour $m = E(X_0)$ et $\sigma^2 = Var(X_0)$.

A partir d'un échantillon X, il est naturel de regarder sur X les valeurs observées des moyenne et variance empiriques pour en inférer les valeurs (ou un encadrement) de ces paramètres pour la population. La loi des grands nombres nous garantira une bonne probabilité pour qu'avec n assez grand, on ne soit pas trop loin des bons résultats.

C'est ainsi que se pose le problème de l'estimation. D'où l'intérêt d'introduire les statistiques T(X) suivantes, définies sur l'échantillon :

$$\bar{X} = \frac{1}{n}\sum X_i \qquad\qquad \Sigma^2 = \frac{1}{n}\sum (X_i - m)^2$$

La 2ème est souvent peu utile, car m est généralement inconnu ; on ne peut donc calculer la valeur observée. On cherche alors à remplacer m par une valeur approchée.

Or $\bar{X}$ est considéré (on le verra plus loin) comme un "bon" estimateur de m. En prenant la valeur observée $\bar{x}$ à la place de celle de m, on obtient la statistique :

$$S'^2 = \frac{1}{n}\sum (X_i - \bar{X})^2 \quad \text{qui en réalité est moins "bonne" que} \quad S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$$

Ces deux statistiques $\bar{X}$ et S^2 font l'objet de nombreux résultats intéressants. On ne va pas les démontrer ici, on se reportera à la bibliographie [7, p.266].

Pour $\bar{X}$

* $E(\bar{X}) = m, \quad Var(\bar{X}) = \sigma^2/n$
* $\bar{X} \to m$, en proba (loi faible des grands nombres)
* $\bar{X} \to m$, p.s. (loi forte)
* $\dfrac{\bar{X} - m}{\sigma/\sqrt{n}} \xrightarrow{\text{loi}} U \in \mathcal{N}(0,1)$ (théorème central-limit)

On utilisera l'approximation en loi $\bar{X} \approx \mathcal{N}(m, \sigma^2/n)$ dès que n>30 ([3]). Cette dernière propriété permet de calculer les probabilités pour que $\bar{X}$ s'écarte de m de plus d'une valeur donnée. Cela permet donc d'établir les fourchettes d'encadrement pour m avec leur fiabilité, à la base de la détermination des intervalles de confiance.

Pour S^2

* $E(S^2) = \sigma^2$ (raison pour la préférer à S'^2, car $E(S'^2) = \sigma^2 - \sigma^2/n$)

* $\mathrm{Var}(S^2) = \frac{\mu_4 - \sigma^4}{n} + \frac{2\sigma^4}{n(n-1)}$, μ_4 moment centré d'ordre 4 de la loi de X_0

* $\mathrm{cov}(\bar{X}, S^2) = \frac{\mu_3}{n}$ (non corrélées si $\mu_3 = 0$, de toute façon asymptotiquement non corrélées)

* $S^2 \xrightarrow{p.s.} \sigma^2$ (loi forte des grandes nombres)

$\frac{S^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}}\sqrt{n} \xrightarrow{\text{loi}} U \in \mathcal{N}(0,1)$ (application du théorème central-limit et autres théorèmes de convergence)

Dans la pratique, avec une erreur uniforme sur les probabilités calculées inférieure à 10^{-3} ([4]), on fait les approximations suivantes :

dès que n > 30, $\bar{X} \approx \mathcal{N}(m, \sigma^2/n)$

dès que n > 50, $S^2 \approx \mathcal{N}(\sigma^2, \frac{\mu_4 - \sigma^4}{n})$.

Cas particulier d'un échantillon gaussien

Si $X_0 \sim \mathcal{N}(m, \sigma^2)$, alors :

* $\bar{X} \sim \mathcal{N}(m, \sigma^2/n)$ exactement pour tout n
* $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ (loi du χ^2 à n-1 degrés de liberté)
* $\bar{X}$ et S^2 sont indépendantes
* $\frac{\bar{X} - m}{S} \sim T_{n-1}$ (T_{n-1} de Student à n-1 degrés de liberté)

Cette dernière statistique présente l'intérêt de ne pas faire intervenir σ.

Mais pour n > 50, $T_{n-1} \sim \mathcal{N}(0,1)$, d'où son utilisation essentiellement pour les "petits" échantillons.

IV- ESTIMATEURS

A) CADRE DE L'ESTIMATION

L'estimation consiste à donner des valeurs approchées aux paramètres d'une population (m, σ,...) à l'aide de l'observation d'un n-échantillon d'une variable parente X_0.

On s'intéresse de préférence aux valeurs d'un caractère représenté par X_0 : sa moyenne m, sa variance σ^2, la proportion p d'objets d'un certain type A dans la population, et on introduit les statistiques

---

3) si la loi de X_0 n'est pas trop dissymétrique, n>50 ou plus pour une loi fortement dissymétrique.

4) loi assez régulière.

$\bar{X}$, S^2 et F (où F est la fréquence du caractère dans l'échantillon) ayant les propriétés d'estimateurs convergents :

$$\bar{X} \xrightarrow{p.s.} m, \quad S^2 \xrightarrow{p.s.} \sigma^2, \quad F \xrightarrow{p.s.} p$$

On a déjà étudié $\bar{X}$ et S^2 ; remarquons que F est un cas particulier de $\bar{X}$:
F est la fréquence empirique de la réalisation par X_0 d'une qualité donnée : situation de Bernoulli,

$$X_o = \begin{cases} 1 & \text{si qualité obtenue} \\ 0 & \text{sinon} \end{cases}, \quad \text{(avec la probabilité p)}$$

Alors $X = (X_1,...,X_n)$ est une "bernoullade" (5) et $F = {}^1/_n \Sigma X_i$ désigne la fréquence du type A dans l'échantillon. Donc $F = \bar{X}$. La loi de F est connue : $n.F \sim B(n,p)$. Le paramètre à estimer est p :

On a $E(X_0) = p$, $var(X_0) = p.q$, $E(F) = p$, $var(F) = \frac{pq}{n}$

la loi de X est donnée par les probabilités des valeurs $x = (x_1,...,x_n)$: $p^{\Sigma x_i}.q^{n-\Sigma x_i}$
la loi de F est : $P(F = f) = P(\Sigma X_i = nf) = \Sigma_x P(X=(x_1,...,x_n) / \Sigma x_i = nf)$

or les x qui vérifient $\Sigma x_i = nf$ sont au nombre de C_n^{nf} parmi les 2^n valeurs possibles,

d'où $P(F = f) = C_n^{nf} p^{nf}.q^{n(1-f)}$

$f_F(f,p) = C_n^{nf} p^{nf}.q^{n(1-f)}$ est la fonction de vraisemblance du modèle qui permet d'estimer une proportion p.

B) QUALITES D'UN ESTIMATEUR

Si T est susceptible d'estimer θ, on espère que la valeur t observée sera "proche de θ".

Il y a deux manières d'être proche :

1- <u>en moyenne</u> : si on réalise de nombreux échantillonnages, les valeurs de t seront réparties autour de θ. Cela se traduit dans le modèle par la propriété : $E_\theta(T) = \theta$; E_θ parce que cette espérance est calculée sur la base de la loi de T qui dépend de θ.

Cette propriété est celle d'un estimateur <u>sans biais</u> (e.s.b., c'est le cas des précédents).

Par exemple $E_\sigma(S'^2) = \sigma^2 - \sigma^2/n$. Le biais de S'^2 est σ^2/n, ce n'est pas rédhibitoire ; par une homothétie, on peut toujours rendre sans biais un estimateur, à condition de savoir calculer son biais.

2- <u>asymptotiquement</u> : plus n est grand (plus on paye cher l'échantillonnage), plus on se rapproche de θ. En probabilité, c'est le sens le plus opératoire, presque sûrement, c'est le plus fort.

Cela se traduit par : $T_n \xrightarrow[n\to\infty]{} \theta$, en proba.

C'est la propriété d'être <u>convergent</u> pour T_n, et pour n assez grand, de minimiser la probabilité : $P_\theta(|T_n - \theta| > \varepsilon)$. Le n assez grand et la valeur de cette probabilité supposent de connaître la loi de T_n.

Les estimateurs précédents sont convergents ; si T_n est sans biais et si $var(T_n) \to 0$, alors T_n est convergent (inégalité de Bienaymé-Tchebychev).

3- <u>Efficacité, limite théorique des performances d'un estimateur</u>

Pour minimiser la probabilité précédente, on cherche les estimateurs les plus précis, qui sont donc les moins dispersés possibles autour de la valeur θ, pour un n donné, ceci en moyenne. Cela sera traduit par la recherche d'un minimum pour l'espérance $E_\theta(|T - \theta|^2)$ (= $var_\theta(T)$ si T est un e.s.b. de θ en dim. 1).

Cette espérance est appelée risque quadratique $R(\theta,T)$ de l'estimateur T.

Lorsque T réalise ce minimum, il est dit "admissible" (il n'est pas forcément sans biais).

Un "bon" estimateur sera alors un e.s.b. convergent de variance minimale.

---

5) c'est-à-dire un n-échantillon de Bernoulli !

Cette variance minimale ne peut être aussi petite que l'on veut, il y a une borne inférieure donnée par un théorème important de la statistique, le théorème de Cramer-Rao-Fréchet-Darmois :

Si $f_X(x,\theta)$ est la fonction de vraisemblance du modèle (existence supposée), avec certaines hypothèses de régularité, en posant :

$$I_n(\theta) = E_\theta\left[\left(\frac{\partial \ln f_X(x,\theta)}{\partial \theta}\right)^2\right]$$ si cette espérance existe (quantité d'information de Fisher),

alors, si T est un e.s.b. de θ défini sur X dont la variance existe, on a :

$$Var_\theta(T) \geq \frac{1}{I_n(\theta)}$$ (borne inférieure de Rao-Cramer).

Si T réalise l'égalité, c'est l'e.s.b. le meilleur, il est dit efficace. C'est le cas de $\bar{X}$, F, et de Σ^2 quand m est connu.

C) RECHERCHE D'UN ESTIMATEUR : METHODE DU MAXIMUM DE VRAISEMBLANCE

On peut penser que pour les exemples que nous avons pris, nous avons les meilleurs estimateurs.

Dans le cas général, il y a des critères pour vérifier qu'un tel estimateur existe et que l'on a le meilleur estimateur (Th. de Rao-Blackwell et Lehmann-Scheffé dans le cas de statistiques exhaustives, c'est-à-dire conservant l'information contenue dans X, et calculable quand le modèle est de type exponentiel) (Saporta [7, p.291]).

Empiriquement, il y a une méthode qui conduit le plus souvent à l'expression d'un bon estimateur : la méthode du maximum de vraisemblance.

Elle consiste à prendre comme estimation de θ la valeur $\hat{\theta}$ qui rend maximale la fonction de vraisemblance $f_X(x,\theta)$.

Cela revient à supposer que l'échantillon observé était le plus "probable" puisque f_X désigne une densité de probabilité. C'est explicite dans le cas discret où f_X est la valeur des probabilités élémentaires, c'est intuitif dans le cas continu où, pour un encadrement donné ε de cette valeur de θ, on a le maximum de la probabilité de s'y trouver lorsque $\hat{\theta}$ réalise le maximum de f_X.

Cela procède d'une hypothèse empirique : l'événement observé est le plus probable (on a plus de chance de le voir que les autres), ce qui est une hypothèse hardie...

Avec cette méthode, la détermination de $\hat{\theta}$ est du domaine mathématique, à partir du problème :

$$\forall\, x \in E,\ f_X(x, \hat{\theta}) = \sup_{\theta \in \Theta} f_X(x, \theta),$$

A chaque $x \in E$ on obtient (peut-être) une valeur pour $\hat{\theta}$ qui est alors une fonction de x, donc une statistique définie sur E , l'estimateur du maximum de vraisemblance ; pratiquement :

- ou on trouve $\hat{\theta}(x)$ facilement, vu la tête de f_X
- ou on a recours à l'analyse et aux régularités de f_X.

Mais en échantillonnage, $f_X(x,\theta) = \prod_{i=1}^{n} f_{X_o}(x_i,\theta)$; or maximiser un produit de termes positifs revient à maximiser son logarithme ; ceci revient à poser que $\hat{\theta}$ satisfait

$\frac{d}{d\theta} \ln f_X(x,\theta) = 0$, en vérifiant qu'on a bien un maximum ; cela s'écrit : $\sum_{i=1}^{n} \frac{d}{d\theta} \ln f_{X_o}(x_i,\theta) = 0$

[équation de vraisemblance] qui est une fonction implicite de θ en x dont la solution est $\hat{\theta}(x)$.

exemple - loi de Bernoulli : estimation de p.

$X_o \sim \mathcal{B}(1,p)$, $f_X(x,p) = p^{\Sigma x_i}.(1-p)^{n-\Sigma x_i}$

équation de vraisemblance : $\frac{d}{dp}\ln f_X(x,p) = \frac{\Sigma x_i}{p} - \frac{n-\Sigma x_i}{1-p} = 0$

solution : $np - \Sigma x_i = 0$, $p = \frac{\Sigma x_i}{n}$, avec $\frac{d^2}{dp^2}\ln f_X(x,p) = -\frac{\Sigma x_i}{p^2} - \frac{n-\Sigma x_i}{(1-p)^2} < 0$.

On a bien un maximum sur]0, 1[, d'où l'estimateur du maximum de vraisemblance : $\hat{p} = \bar{X}$.

exemple - loi normale. Dans ce cas, θ est à plusieurs dimensions : $\theta = (m,\sigma^2)$, on obtient par le même principe un système d'équations aux dérivées partielles. Traitons l'exemple :

$X_o \sim \mathcal{N}(m, \sigma^2)$, $f_X(x,\theta) = \frac{1}{(2\pi)^{n/2}\sigma^n}.e^{-\frac{1}{2}\Sigma\left(\frac{x_i-m}{\sigma}\right)^2}$, avec un extremum local pour f_X si

$$\frac{\partial}{\partial m}\ln f_X(x,\theta) = 0 \quad \text{et} \quad \frac{\partial}{\partial(\sigma^2)}\ln f_X(x,\theta) = 0$$

La première équation donne : $\Sigma\frac{x_i - m}{\sigma^2} = 0$, d'où $m = \frac{\Sigma x_i}{n}$ et $\hat{m} = \bar{X}$.

La deuxième donne : $-\frac{n}{2}\times\frac{1}{\sigma^2} - \frac{1}{2}\Sigma\frac{(x_i-m)^2}{\sigma^4} = 0$, d'où $\sigma^2 = 1/n\ \Sigma(x_i - \bar{x})^2$, et $\hat{\sigma^2} = S'^2$ (biaisé).

Cela, si on vérifie qu'on a bien un maximum (local), condition donnée pour une fonction de 2 variables par $rt - s^2 > 0$. Ici, le calcul des dérivées partielles secondes avec m et σ^2 et l'expression donne bien

$$rt - s^2 = -\frac{n^2}{2\sigma^6} + \frac{n^2}{\sigma^6} > 0$$

V- ESTIMATION PAR INTERVALLE

A) POSITION DU PROBLEME

Fournir une valeur numérique pour un paramètre à estimer (estimation ponctuelle) n'est pas satisfaisant en pratique, où il faut pouvoir quantifier les risques pris en retenant cette valeur. En effet, cela ne fournit ni la marge d'erreur : la précision de la donnée numérique, ni la probabilité que l'échantillon prélevé conduise effectivement à une "bonne" valeur : la fiabilité du procédé (on pourrait être très malchanceux avec l'échantillon !).

D'où le problème de l'estimation par intervalle ainsi posé pour le paramètre θ en dimension 1 :

Trouver un intervalle [a(X) ; b(X)] dont les bornes sont aléatoires puisque déterminées par l'échantillon X et tel que $P_\theta(\theta \in [a(X) ; b(X)]) \geq 1 - \alpha$.

* l'échantillon étant alors prélevé, son observation x fournit l'intervalle réel [a(x) ; b(x)] appelé "fourchette" pour l'estimation de θ, sans qu'on puisse en toute certitude dire si la valeur réelle θ_0 du paramètre θ pour la population $\mathcal{P}$ est entre a(x) et b(x).
* $1 - \alpha$ est appelé le niveau de confiance (par ex. 0,95 ou 95%), indice de fiabilité du résultat et α est le risque (de tomber à côté !).

* la probabilité P_θ qu'on cherchera à rendre la plus voisine de 1- α (pour des raisons d'économie), est calculable si on connaît la loi de X, ou plutôt de T statistique servant à déterminer les bornes a(X) et b(X).
* l'intervalle est d'autant plus étroit que T est moins dispersée, d'où l'intérêt d'avoir des estimateurs sans biais le plus efficaces possibles, de loi connue et convergents "rapidement" pour minimiser la taille n de l'échantillon nécessaire pour réaliser le niveau 1- α.
* élargir l'intervalle de confiance (perdre en précision), c'est augmenter P_θ et donc se permettre d'atteindre le niveau de confiance, c'est en fait diminuer le risque α (gagner en fiabilité). On sera toujours à la recherche du compromis entre précision et fiabilité.
* la valeur de α, donnée a priori (à moins que ce soit n), est déterminée par les coûts que représentent les fausses informations (cas où θ_0 n'est pas dans la fourchette). Ces coûts sont calculables sur un grand nombre d'estimations liées à une pratique industrielle et commerciale. Sans les contraintes financières, la détermination d'un intervalle de confiance n'a d'intérêt que qualitatif lorsque des valeurs standards de α sont données (en médecine ou sciences humaines par exemple), on prend alors souvent $\alpha = 0{,}05$.

B) PRINCIPE DE LA DETERMINATION D'UN INTERVALLE DE CONFIANCE

La solution du problème précédent où on doit déterminer a(X) et b(X) n'est pas unique : 1 relation pour 2 inconnues. Les conditions du problème restreignent cette indétermination.

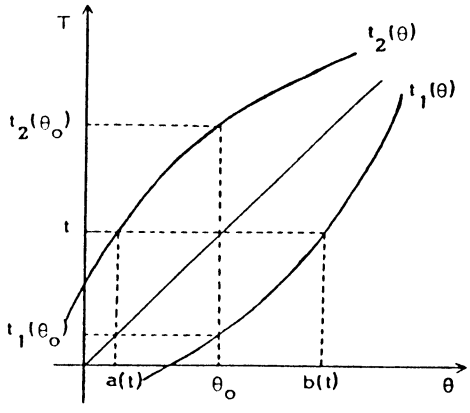
On peut chercher • un maximum de confiance : $P_\theta(\theta \le M(X)) = 1 - \alpha$

• ou un minimum : $P_\theta(m(X) \le \theta) = 1 - \alpha$

• ou penser qu'il n'y a pas de raison pour que le risque soit plus porté à droite qu'à gauche de l'intervalle et équilibrer ce dernier par les conditions : $P_\theta(\theta \le a(X)) = \alpha/2$ et $P_\theta(\theta \ge b(X)) = \alpha/2$.

Plaçons nous dans cette situation.

Si on a un "bon" estimateur T de θ et si, pour chaque valeur de θ, on trouve deux réels t_1 et t_2 tels que $P_\theta(t_1 < T < t_2) = 1 - \alpha$, on peut penser que la valeur θ_0 de θ est proche de la valeur observée t de T et considérer que si $t \in [t_1, t_2]$ cette valeur est possible pour θ_0. Cela se traduit par le graphique



d'où la détermination
$a(t) = t_2^{-1}(t)$
$b(t) = t_1^{-1}(t)$
si c'est possible

alors $t_1(\theta_0) < T(X) < t_2(\theta_0)$
équivaut à
$t_2^{-1}(T) < \theta_0 < t_1^{-1}(T)$
de probabilité 1- α

En théorie, il faudrait admettre que les fonctions $t_1(\theta)$ et $t_2(\theta)$ (non uniques) sont croissantes et inversibles localement. On préfère résoudre directement ce problème d'inversion d'inégalités dans chaque cas, comme nous allons le voir sur un exemple.

choix d'une statistique de décision.

En réalité, l'estimateur T ne contient pas explicitement le paramètre θ qui intervient dans le calcul des bornes t_1 et t_2 par l'intermédiaire de la loi P_θ, inconnue.

On transforme alors cette statistique en une statistique contenant θ mais dont la loi n'en dépend plus et est bien connue, permettant de calculer la probabilité demandée. Cette dernière statistique sera la variable de confiance.

Exemple : estimation de la moyenne d'une loi normale avec écart-type connu.

$X_0 \sim \mathcal{N}(m, \sigma^2)$; $\bar{X} \sim \mathcal{N}(m, \sigma^2/n)$ est le meilleur estimateur de m, mais sa loi contient m.

On prend alors $\dfrac{X - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ qui sera la variable de confiance adéquate.

C) MISE EN PRATIQUE DE LA RECHERCHE DE L'INTERVALLE DE CONFIANCE

Il faudra l'adapter à chaque cas d'espèce, d'où un ensemble de résultats particuliers donnés dans les manuels [8, p.149] ou [7, p.304].

Reprenons l'exemple précédent ; on part de la condition

$P_m(a(X) < m < b(X)) = 1 - \alpha$ équivalente par transformations affines à une condition du type :

$P_m(-u_{\alpha/2} < U < u_{\alpha/2}) = 1 - \alpha$, où $u_{\alpha/2} > 0$ est le fractile d'ordre $\alpha/2$ de la loi normale centrée réduite, défini par $P[U > u_{\alpha/2}] = \alpha/2$ (lu dans la table : $\alpha = 0{,}05$ donne $u_{\alpha/2} = 1{,}96$).

On a bien $-u_{\alpha/2} < U < u_{\alpha/2} \Leftrightarrow -u_{\alpha/2}\,\sigma/\sqrt{n} < \bar{X} - m < u_{\alpha/2}\,\sigma/\sqrt{n}$

$\Leftrightarrow \bar{X} - u_{\alpha/2}\,\sigma/\sqrt{n} < m < \bar{X} + u_{\alpha/2}\,\sigma/\sqrt{n}$

d'où l'intervalle de confiance équilibré :

$P[\bar{X} - u_{\alpha/2}\,\sigma/\sqrt{n} < m < \bar{X} + u_{\alpha/2}\,\sigma/\sqrt{n}] = 1 - \alpha$

et la fourchette obtenue avec l'observation $\bar{x}$ de $\bar{X}$: $(\bar{x} - u_{\alpha/2}\,\sigma/\sqrt{n}\ ;\ \bar{x} + u_{\alpha/2}\,\sigma/\sqrt{n})$,

avec les remarques d'usage :

- La fiabilité $1 - \alpha$ influe sur la valeur $u_{\alpha/2}$, ici 1,96 si $\alpha = 0{,}05$; 1,65 si $\alpha = 0{,}1$, donc peu influente devant les variations de n (l'écart de la fourchette dû à un choix de fiabilité raisonnable peut varier du simple au double) ;

- l'écart de la fourchette est proportionnel à σ, écart-type de la population, qui trouve ici son sens concret. On retrouve le fait que la fourchette est d'autant resserrée que la dispersion de X_0 est faible ;

- l'écart (la précision) est inversement proportionnel à $\sqrt{n}$: il faut multiplier par 4 le nombre d'observations pour réduire de moitié l'intervalle de confiance. Ce sera le cas général de l'estimation d'une moyenne du fait que $\mathrm{Var}(\bar{X}) = \sigma^2/n$;

- numériquement, si X_0 est une mesure en cm autour de 20 cm et $\sigma = 1$ cm (les 2/3 des mesures tombent entre 19 et 21 cm), avec $\alpha = 0{,}05$ et $n = 49$, l'intervalle observé est pour m : [19,7 ; 20,3].

Dans cet exemple, avec l'hypothèse forte $X_0 \sim \mathcal{N}(m, \sigma^2)$, on voit qu'un échantillon modeste de taille 50, donne un intervalle appréciable de longueur 0,6 pour les données numériques choisies.

Dans l'estimation d'une proportion où l'on ne connaît pas la loi de la variable qui conduit à attribuer telle ou telle qualité au caractère, on verra que la taille de l'échantillon doit être bien plus grande pour estimer avec une bonne fiabilité cette proportion (n = 1000 dans les sondages, cf. l'atelier).

D) FORMULATION DE LA CONCLUSION

Revenons à l'exemple.

Mathématiquement, la réponse est : l'intervalle de confiance pour estimer m au niveau de confiance $1 - \alpha$ est : $[\bar{X} - u_{\alpha/2}\,\sigma/\sqrt{n}\ ;\ \bar{X} + u_{\alpha/2}\,\sigma/\sqrt{n}]$.

Cette formulation ne peut satisfaire l'utilisateur qui doit prendre des décisions. De plus il aimerait avoir de bonnes valeurs numériques justifiées par l'échantillon qu'il a payé assez cher.

On retombe sur un problème épistémologique :

- ou on ne fera ce prélèvement qu'une seule fois, et mon penchant fréquentiste a de la difficulté à donner du sens opératoire à la notion de probabilité (que θ soit effectivement dans l'intervalle retenu), car du point de vue numérique, θ est, ou n'est pas, dans l'intervalle observé ; il n'y a plus, après cette observation, d'expérience aléatoire, donc de probabilité ;

- ou il fait partie d'un contrôle régulier, répété un assez grand nombre de fois, et je préfère formuler le résultat en les termes suivants :
au niveau de confiance $1 - \alpha = 0{,}95$ (par exemple), ma méthode mathématique conduit à un résultat

$(m \in [\ \bar{x} - u_{\alpha/2}\ \sigma/\sqrt{n}\ ;\ \bar{x} + u_{\alpha/2}\ \sigma/\sqrt{n}\ ])$ vrai en moyenne 95 fois sur 100 (mais je ne sais pas quand je me trompe). Vous pourrez donc me faire relativement confiance.

Pour vous, utilisateur, vous pouvez savoir que si vous utilisez un grand nombre de fois une telle estimation de m (tous les mois par exemple) dans votre contrôle de production, sachez que 5 fois sur 100 en moyenne l'intervalle observé (la fourchette) ne contiendra pas la vraie valeur de m. Vous pourrez alors évaluer les coûts que cela représente pour vous, prendre vos dispositions, et revoir peut-être le niveau de confiance que vous m'avez donné, en fonction de la taille n de l'échantillon que vous acceptez de payer.

BIBLIOGRAPHIE

[1] LAPLACE Pierre Simon : *Essai philosophique sur les probabilités* (1825) Editions Ch. BOURGEOIS, 1986.

[2] POINCARE Henri : *La Science et l'hypothèse* (1902) Editions CHAMPS-FLAMMARION, 1968.

[3] POINCARE Henri : *Calcul des probabilités* (1912), réédition J. GABAY, 1987.

[4] EKELAND Ivar : *Au hasard*, Editions du SEUIL, 1990.

[5] BERNOULLI Jacques : *Ars Conjectandi* (1713), traduction de N. Meusnier, Brochure IREM de Rouen, 1987.

[6] INTER-IREM : *Histoire et épistémologie : Maths au fil des âges* Editions GAUTHIER-VILLARS, 1987.

[7] SAPORTA Gilbert : *Probabilités, Analyse des données et Statistique* Editions TECHNIP, 1990.

[8] BIGOT Bernard et VERLANT Bernard : *Mathématiques, statistiques et probabilités*, cours de BTS Editions FOUCHER, 1990.

[9] HENRY Michel et Annie : L'enseignement des probabilités dans le nouveau programme de Première, in *Repères-IREM*, n°6, 1992.

[10] IREM de BESANÇON : *L'enseignement des statistiques et des probabilités en STS* Brochure IREM de Besançon, 1990.

Atelier

Pratique de la recherche d'un intervalle de confiance pour l'estimation d'une proportion p. Application aux sondages

---

Modelisation

Dans la population P, une proportion p des individus possède le caractère observé. On cherche un intervalle de confiance pour p.
On fait un tirage avec remise d'un n-échantillon (ou sans remise si la population est assez vaste et si l'échantillonnage ne modifie pas p).

Soit F la proportion (ou fréquence) trouvée dans le n-échantillon des individus ayant le caractère examiné. F est un estimateur sans biais, efficace de p (exercice 33).

On a $nF \sim \mathcal{B}(n, p)$.

L'expression de cette loi est trop compliquée pour donner un intervalle de confiance explicite bâti sur cette variable.

a) Petits échantillons :

On utilise une table de la loi binômiale pour déterminer pour différents p les valeurs $k_1(p)$ et $k_2(p)$ telles que :

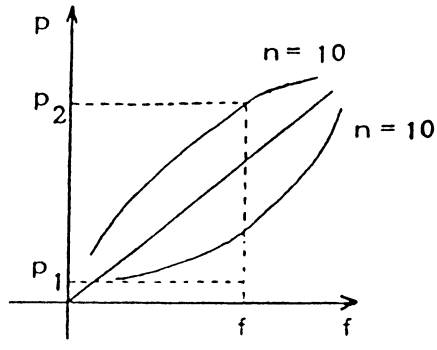
$$P\left(k_1 < nF < k_2\right) = \sum_{k=k_1(p)}^{k_2(p)} C_n^k \, p^k (1-p)^{n-k} = 1-\alpha$$

(avec par exemple $\sum_{k=0}^{k_1} C_n^k \, p^k (1-p)^{n-k} = \frac{\alpha}{2}$).

L'intervalle de confiance pour p (ou plutôt une de ses réalisations) sera déterminé par l'ensemble des p tels que si nf est l'observation issue de l'échantillon, on a $k_1(p) < nf < k_2(p)$ selon la présentation théorique du début.

Pour éviter de nombreux calculs fastidieux, on utilise des abaques de la loi binômiale construites à cet effet.

Une telle abaque est un réseau de courbes. Chaque courbe corresponc à une taille d'échantillon. Elle donne les bornes p_1, p_2 de l'intervalle de confiance pour p en fonction de l'observation f, selon le schéma suivant :



b) Grands échantillons (tirage avec remise ou ne modifiant pas sensiblement p) :

Une nouvelle application du théorème central limite montre que l'on a l'approximation :

$$nF \approx \eta(np, \sqrt{np(1-p)})$$

d'où

$$F \approx \eta\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \text{ et } \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \simeq U \sim \eta(0,1)$$

On obtient, comme pour le 1er paragraphe, l'intervalle de confiance

$$(*) \qquad F - u_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} < p < F + u_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

mais p figure dans les bornes et il faut résoudre en p cette double inégalité. On a trois solutions :

α - Utiliser un agrandissement de l'intervalle de confiance par la majoration $p(1-p) < \frac{1}{4}$ (car $0 < p < 1$) d'où l'intervalle

$$\boxed{F - \frac{u_{\alpha/2}}{2\sqrt{n}} < p < F + \frac{u_{\alpha/2}}{2\sqrt{n}}}$$

qui suppose en fait que p est voisin de $\frac{1}{2}$ (mais on sait que pour que l'approximation normale soit valable, il ne faut pas que p soit trop proche de 0 ou de 1). On a alors un intervalle de confiance de niveau supérieur à $1-\alpha$, mais on ne connaît pas le niveau exact.

β - Faire une résolution graphique de (*)

Les bornes de l'encadrement (*) de la variable de confiance sont

$$f = p \pm u_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

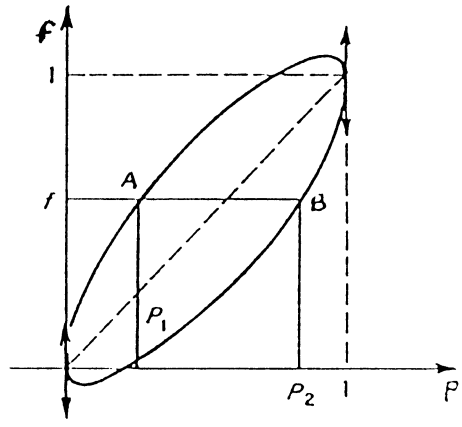
d'où

$$(f-p)^2 = u^2_{\frac{\alpha}{2}} \frac{p(1-p)}{n}$$

d'où l'équation

$$f^2 + p^2\left(1 + \frac{u^2_{\alpha/2}}{n}\right) - 2pf - \frac{u^2_{\alpha/2}\, p}{n} = 0$$

d'une ellipse passant par l'origine et le point (1, 1)



Les points intérieurs à l'ellipse vérifient les inégalités *.

Pour chaque observation f de F, on obtient donc (comme dans le cas des petits échantillons) un couple $p_1(f)$ et $p_2(f)$ et l'intervalle [A, B] représente aussi bien l'évènement $F - u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < F + u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ que $p_1(F) < p < p_2(F)$ de probabilité $1-\alpha$.

Lorsque n (grand) varie, on obtient aussi une abaque constituée d'ellipses dont l'utilisation est analogue à celle des petits échantillons.

Lorsque $n \to +\infty$, les ellipses se rétrécissent et tendent vers la diagonale.

γ - Utiliser F comme estimation ponctuelle de p, on obtient alors l'intervalle de confiance

$$\boxed{F - u_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)}{n}} < p < F + u_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)}{n}}}$$

de niveau $1-\alpha$.

[Ceci est justifié par le calcul suivant : résoudre en p l'équation de l'ellipse donnée

$$p = \frac{\left(2f + \frac{u^2_{\alpha/2}}{n}\right) \pm \sqrt{\frac{u^4_{\alpha/2}}{n^2} + 4f\,\frac{u^2_{\alpha/2}}{n} - 4f^2\,\frac{u^2_{\alpha/2}}{n}}}{2\left(1 + \frac{u^2_{\alpha/2}}{n}\right)}$$

Comme n est grand, on prend plutôt un équivalent, ce qui donne

$$\frac{2f + \frac{u^2_{\alpha/2}}{n}}{2\left(1 + \frac{u^2_{\alpha/2}}{n}\right)} \sim f$$

et

$$\frac{\sqrt{\frac{u^4_{\alpha/2}}{n^2} + 4f\,\frac{u^2_{\alpha/2}}{n} - 4f^2\,\frac{u^2_{\alpha/2}}{n}}}{2\left(1 + \frac{u^2_{\alpha/2}}{n}\right)} \sim \frac{u_{\alpha/2}}{\sqrt{n}}\sqrt{f(1-f)}$$

d'où $p \sim f \pm u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}$.

Exercice

1) Un encadrement de confiance de p est exigé à $\pm$ 0,01 ("fourchette" à 2 % pour un sondage par exemple) avec un niveau $1-\alpha = 95$ %.
On sait que la valeur f observée sera voisine de 0,5 (cas le plus défavorable où $\sqrt{f(1-f)}$ atteint son maximum $\frac{1}{4}$).
Déterminer la taille de l'échantillon nécessaire (Rep. : 9600).

2) Avec un échantillon de 1000 personnes et une fourchette à 2 %, quel est le niveau de confiance du sondage ? (Rep. : 0,46).

c) Grands échantillons exhaustifs (taille de la population = N) :

F est encore un estimateur sans biais de p, on a $\text{var } F = \frac{p(1-p)}{n}\,\frac{N-n}{N-1}$.

On est alors ramené à l'intervalle de confiance précédent, de la forme

$$F - u_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)(N-n)}{n(N-1)}} < p < F + u_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)(N-n)}{n(N-1)}}$$

Exercice

Construire un intervalle de confiance à 10 % pour le paramètre p de Bernoulli si une observation d'un échantillon de taille 50 donne $\Sigma x_i = 15$. (Utiliser l'approximation normale et comparer les résultats dans les deux cas où p(1-p) est remplacé par $\frac{1}{4}$ ou par son estimation tirée de l'échantillon).

Exercice

1) Lancer une pièce 50 fois et compter le nombre de piles. Déterminer alors un intervalle de confiance pour p à 90 %.

2) Lancer un dé 15 fois et compter le nombre de 6. Déterminer un intervalle de confiance pour la probabilité d'obtenir 6 avec ce dé (utiliser une abaque). Votre dé est-il pipé ?

Exercice

A la veille d'une consultation électorale, on a interrogé 100 électeurs pris au hasard. 64 d'entre eux se sont déclarés favorables au candidat z. Entre quelles limites, au moment du sondage, au niveau 0, 95, la proportion du corps électoral favorable à z se situe-t-elle ?

Exercice

Un médecin désire estimer la proportion des cas qui seront guéris par un nouveau traitement.

a) A combien de patients doit-il appliquer le traitement avant de pouvoir conclure, s'il veut que son estimateur ait un écart type de 0, 005 et s'il pense que le traitement guérira à peu près 75 % des malades.

b) Lors de l'expérience pratiquée en a), quelle est approximativement la probabilité pour que l'estimateur excède 0, 8 alors que, en réalité, le traitement ne soigne que 60 % des cas ?

Voici quelques éléments bibliographiques pour les applications aux sondages:

HENRY Michel: Eléments sur les sondages, cours de statistiques, Maîtrise SMI, Besançon, 1983.

KOSMANEK Edith: Sondages stratifiés, article de "Quadrature n°6, Septembre 1990.

GENET, PUPION et REPUSSARD: Probabilités, statistiques et sondages, cours et exercices corrigés, Vuibert 1974.

SAPORTA Gilbert: Probabilités, analyse des données et statistiques, éditions Technip, 1992.

DROESBEKE Jean-Jacques et TASSI Philippe: Histoire de la statistique, col. "que sais-je", PUF 1990.

Cet abaque est extraite du Manuel de Gilbert SAPORTA: Probabilités, analyse des données et statistiques, éditions Technip, 1992.

Table 3 bis

Abaque donnant en fonction de f
l'intervalle de confiance à 0,95 ($p_{0,025}$ à $p_{0,975}$)

f fréquence observée (en p.100) sur un échantillon d'effectif n
p proportion (en p.100) dans la population échantillonnée