université pierre et marie curie
Master 2 Probabilités et Finance

année universitaire 2015/2016

Quelques Éléments de Statistique Mathématique

Lucien Birgé – Septembre 2015

Chapitre I

Introduction aux problèmes statistiques

1 Notion d'expérience statistique

1.1 Probabilités versus Statistique

Pour illustrer notre propos, partons de l'exemple du jeu de pile ou face qui est modélisé en probabilités par une suite de variables indépendantes $X_1, \ldots, X_n$ de même loi de Bernoulli de paramètre $1/2$ notée $\mathcal{B}(1, 1/2)$, ou de celui de la suite des apparitions du zéro à la roulette que l'on peut modéliser, de manière analogue, par une suite de variables de Bernoulli indépendantes de paramètre $1/37$ soit de loi $\mathcal{B}(1, 1/37)$. De manière plus générale, on peut considérer une suite de variables aléatoires i.i.d. $Y_1, \ldots, Y_n$ à valeurs dans un certain espace probabilisé $(E, \mathcal{E})$, un événement $A \in \mathcal{E}$ et fixer $X_i = \mathbb{1}_A(Y_i)$. On obtient alors une autre *suite d'essais de Bernoulli indépendants* (suite de variables aléatoires i.i.d. à valeurs dans $\{0; 1\}$) de paramètre $p = \mathbb{P}[Y_i \in A] = \mathbb{P}[X_i = 1] \in [0; 1]$, donc de loi $\mathcal{B}(1, p)$. On omettra le plus souvent de préciser "indépendants".

En Théorie des Probabilités, on suppose que p est connu, on s'intéresse au comportement de la suite des X_i et l'on répond à la question suivante : "quelle est la probabilité que, dans la suite des X_i, on trouve k fois la valeur 1 ?" ou, de manière équivalente, "que vaut $\mathbb{P}[S_n = k]$ si $S_n = \sum_{i=1}^n X_i$?" On sait que S_n suit une loi binomiale $\mathcal{B}(n, p)$ et que $\mathbb{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$ pour $0 \leq k \leq n$, que $\mathbb{E}[S_n] = np$ et $\mathrm{Var}(S_n) = np(1-p)$. On démontre aussi la Loi des Grands Nombres (LGN) : "$S_n/n \to_{n\to+\infty} p$ p.s." et le Théorème Limite Central (TLC) : "$n^{-1/2}(S_n - np) \rightsquigarrow \mathcal{N}(0, p(1-p))$ quand $n \to +\infty$", où $\rightsquigarrow$ désigne la convergence en loi, etc.

En Statistique, on raisonne en sens inverse. On ne connaît pas p mais on observe une réalisation $x_1, \ldots, x_n$ avec $x_i = X_i(\omega)$ de la suite d'essais de Bernoulli de paramètre p et l'on utilise ces observations pour tirer des conclusions (nécessairement aléatoires puisque dépendant de ω) sur la valeur inconnue de p.

1.2 Quelques exemples très simples

Exemple 1 Supposons que l'on veuille tester un nouveau médicament et évaluer son efficacité p qui représente la probabilité qu'il a d'être efficace. On administre le médicament à n personnes malades et l'on observe pour chacune d'elles s'il est efficace ou non. En première approximation, cela donne une suite d'essais de Bernoulli de paramètre p. On peut faire le même raisonnement pour tester la toxicité d'une nouvelle molécule, sauf que

là, les essais ne se feront pas sur des malades.

Exemple 2 Pour faire une enquête d'opinion, on interroge n personnes pour savoir si elles s'intéressent ou non à tel produit. Là encore, on peut admettre, en première approximation, qu'il s'agit d'une suite d'essais de Bernoulli de paramètre p, probabilité qu'une personne prise au hasard s'intéresse au produit.

Exemple 3 Même raisonnement pour un sondage électoral avec deux candidats, ou un referendum avec une réponse par oui ou non, etc.

1.3 Modèles statistiques

Quand on fait de la Statistique, on part en général d'un problème concret lié à un phénomène aléatoire et l'on suppose que ce que l'on observe, disons $\boldsymbol{X}$, est le résultat d'une expérience aléatoire. Ici $\boldsymbol{X}$ peut être une suite de variables ou de vecteurs aléatoires, la trajectoire d'un processus, etc. Construire un modèle statistique, c'est se donner pour $\boldsymbol{X}$ une famille de lois possibles : $\{P_\theta, \theta \in \Theta\}$ indexée par un paramètre θ. La famille de lois contient l'information que l'on a sur le phénomène : indépendance, équidistribution, forme de la loi. Par exemple, on observe $\boldsymbol{X} = (X_1, \ldots, X_n)$ formant une suite d'essais de Bernoulli de paramètre θ inconnu. Ceci définit parfaitement la famille des lois P_θ comme une famille de lois produits.

1.3.1 Modèles approchés

Il faut toujours avoir à l'esprit que tout modèle statistique est un modèle "approché", c'est à dire une approximation plus ou moins bonne de la réalité. C'est évidemment vrai dans les exemples vus plus haut. Pour les médicaments, il y a bien des chances, même si l'on a sélectionné des malades analogues, qu'il n'y ait pas équidistribution. Dans une enquête d'opinion, ou pour un sondage on a le même problème car la population n'est pas homogène vis à vis des questions. De plus on n'interroge jamais deux fois la même personne, ce qui pourrait arriver si on choisissait vraiment les gens "au hasard". On fait souvent, délibérément, des approximations de la réalité pour simplifier les choses. Une approximation très courante, par exemple, est de supposer que certaines variables sont gaussiennes standard, c'est à dire de loi $\mathcal{N}(0,1)$. Mais la loi normale, de densité $(2\pi)^{-1/2}\exp\left[-x^2/2\right]$ est une loi à support sur $\mathbb{R}$ tout entier alors que les phénomènes que l'on mesure sont essentiellement bornés. D'un autre côté, un calcul facile montre que si l'on a n gaussiennes standard indépendantes $X_1, \ldots, X_n$,

$$\mathbb{P}\left[\sup_n |X_i| > t\right] \le 2\mathbb{P}\left[\sup_n X_i > t\right] = 2\left[1 - (\mathbb{P}[X_1 \le t])^n\right] = 2\left[1 - (1 - \mathbb{P}[X_1 > t])^n\right].$$

Or on peut montrer l'inégalité très utile, valable pour $t > 0$,

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{t^2}{2}\right] < \mathbb{P}[X_1 > t] < \frac{1}{t}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{t^2}{2}\right]. \qquad \text{(I.1.1)}$$

Si $t = 6$, $\mathbb{P}[X_1 > t]$ est inférieure à 10^{-9} et $\mathbb{P}\left[\sup_n |X_i| > t\right]$ est négligeable si $n \le 10^6$. Si $t = 8$, $\mathbb{P}[X_1 > t]$ est inférieure à 6×10^{-16} et $\mathbb{P}\left[\sup_n |X_i| > t\right]$ est négligeable si $n \le 10^{12}$. Pour tout usage pratique, une gaussienne standard est indiscernable d'une variable à support dans $[-10; 10]$.

1.3.2 Expériences et modèles statistiques

Le travail du statisticien comporte deux phases. Une phase dite de modélisation : on considère un problème réel que l'on modélise sous forme mathématique en supposant que l'observation $\boldsymbol{X}$ est un objet (variable, vecteur, processus) aléatoire dont la loi (inconnue) appartient à une certaine famille de lois que l'on précise et qui dépendent d'un certain nombre de paramètres. Cette phase n'est pas facile car elle dépend d'un phénomène réel et donc suppose des informations sur ce phénomène. Une telle modélisation repose sur la connaissance du domaine étudié (physique, économie, finance, ...). Nous ne nous occuperons pas de cela ici, même si c'est très important, parce que cela relève en grande partie du domaine d'application considéré. Tout au plus verrons-nous comment tester si un modèle semble correct.

Définition 1 *Une* expérience statistique *est la donnée d'un objet aléatoire $\boldsymbol{X}$ à valeurs dans un espace mesurable $(E, \mathcal{E})$ et d'une famille de lois $\{P_\theta, \theta \in \Theta\}$ sur cet espace. On suppose que la vraie loi de $\boldsymbol{X}$ appartient à cette famille. On dira que $\{P_\theta, \theta \in \Theta\}$ constitue un* modèle statistique *pour la loi de $\boldsymbol{X}$.*

On rappelle que $\boldsymbol{X}$ est une application mesurable de $(\Omega, \mathcal{A})$ dans $(E, \mathcal{E})$ et que la loi $P_{\boldsymbol{X}}$ de $\boldsymbol{X}$ est l'image par $\boldsymbol{X}$ d'une probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{A})$. On suppose donc qu'il existe une famille de probabilités $\{\mathbb{P}_\theta, \theta \in \Theta\}$ sur $(\Omega, \mathcal{A})$ qui induit la famille des lois P_θ possibles pour $\boldsymbol{X}$ et l'on en tiendra compte pour les notations :

$$P_\theta(A) = \mathbb{P}_\theta[\boldsymbol{X} \in A] = \mathbb{P}_\theta[\{\omega \in \Omega \,|\, \boldsymbol{X}(\omega) \in A\}] \quad \text{pour tout } A \in \mathcal{E}.$$

Plus généralement, lorsque l'on travaillera avec plusieurs lois simultanément, on précisera dans les espérances la loi de la variable que l'on utilise. Quand on note $\mathbb{E}_Q[f(Y)]$, on suppose que Y est de loi Q donc $\mathbb{E}_Q[f(Y)] = \int f(y)\, dQ(y)$. On abrègera $\mathbb{E}_{P_\theta}[f(\boldsymbol{X})]$ (espérance de $f(\boldsymbol{X})$ quand $\boldsymbol{X}$ a la loi P_θ) en $\mathbb{E}_\theta[f(\boldsymbol{X})]$.

Pour en revenir à l'exemple basique des essais de Bernoulli indépendants et de même loi, l'espace $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$ est, comme toujours, une abstraction, mais $(E, \mathcal{E}) = (\{0;1\}^n, \mathcal{P}(\{0;1\}^n)$, $\boldsymbol{X} = (X_1, \ldots, X_n)$ et

$$\mathbb{P}_\theta[X_i = \delta_i, i = 1, \ldots, n] = P_\theta[(\delta_1, \ldots, \delta_n)] = \theta^{\sum_{i=1}^n \delta_i}(1-\theta)^{n-\sum_{i=1}^n \delta_i} \quad \forall \delta \in \{0;1\}^n.$$

Ici $\Theta = [0;1]$ ou un de ses sous-ensembles.

Un modèle très voisin est le modèle binomial. Le plus souvent lorsque l'on compte les "face" à pile ou face, les zéros à la roulette ou lorsque l'on fait une enquête d'opinion ou un sondage, on ne regarde pas l'ordre dans lequel sont apparus les 0 et les 1 des essais de Bernoulli mais on résume tout cela par une seule variable, $N = \sum_{i=1}^n X_i$ à valeur dans $\{0;1;\ldots;n\}$ et de loi binomiale $\mathcal{B}(n,\theta)$ lorsque les X_i sont de paramètre θ. Ceci donne un nouveau modèle statistique qui est essentiellement équivalent au précédent pour la raison suivante. La loi jointe des X_i, conditionnellement à N, se calcule facilement :

$$\mathbb{P}_\theta[X_i = \delta_i, i = 1, \ldots, n \,|\, N = k] = 0 \quad \text{si } \sum_{i=1}^n \delta_i \neq k$$

et si $\sum_{i=1}^n \delta_i = k$, alors

$$\mathbb{P}_\theta[X_i = \delta_i, i = 1, \ldots, n \,|\, N = k] = \frac{\theta^k(1-\theta)^{n-k}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} = \binom{n}{k}^{-1},$$

quel que soit θ. Autrement dit, la loi conditionnelle des X_i sachant N est indépendante de θ. Ceci implique que, si l'on dispose de N, on peut, sans connaître θ, reconstruire des variables aléatoires de même loi jointe que les X_i (par un tirage aléatoire uniforme de la répartiton des N "uns" parmi les n positions possibles). Donc la connaissance des X_i n'apporte aucune information supplémentaire sur θ par rapport à celle de N. Toute l'information sur θ est contenue dans la loi de N. On dit alors que N est une *statistique exhaustive*.

Définition 2 *Étant donné une expérience statistique* $\{P_\theta, \theta \in \Theta\}$*, d'observation* $\boldsymbol{X}$*, on appelle* statistique *tout objet aléatoire (variable, vecteur, ...) fonction de* $\boldsymbol{X}$ *et éventuellement de paramètres* connus *mais indépendant de* θ*.*

Autrement dit, une statistique est une quantité que l'on peut effectivement calculer au vu des observations puisqu'elle ne dépend que de quantités connues.

1.4 Divers types de modèles

On distingue classiquement divers types de modèles statistiques en fonction de la "taille" de l'espace des paramètres. Il y a les modèles dits *paramétriques* pour lesquels Θ est une partie de $\mathbb{R}$ ou, plus généralement, d'un espace euclidien $\mathbb{R}^k$. Le modèle binomial en fait partie, mais aussi le modèle gaussien réel à deux paramètres, c'est à dire le modèle où l'on observe n v.a.r. i.i.d. $X_1, \ldots, X_n$ de loi normale $\mathcal{N}\left(\mu, \sigma^2\right)$ où les paramètres μ et σ^2 sont inconnus. Ces deux exemples sont des cas de modèles *dominés*, c'est à dire pour lesquels il existe une mesure de référence μ telle que toutes les lois P_θ, $\theta \in \Theta$, sont absolument continues par rapport à μ. Le modèle peut alors être défini par la famille des densités $f_\theta = dP_\theta/d\mu$, $\theta \in \Theta$. C'est le cas des lois précédemment évoquées, les binomiales, dominées par la mesure de comptage, ou les gaussiennes, dominées par la mesure de Lebesgue. Les modèles paramétriques les plus classiques sont dominés et même, le plus souvent, directement définis par une mesure de référence μ et une famille de densités $\{f_\theta, \theta \in \Theta\}$ par rapport à μ. Mais on peut aisément fabriquer des modèles paramétriques non dominés comme $\{\delta_\theta, \theta \in \mathbb{R}\}$, où δ_x désigne la masse de Dirac au point x.

On peut considérer des modèles plus complexes où l'espace Θ des paramètres est une partie d'un espace fonctionnel. Par exemple les X_i sont i.i.d. sur $[0;1]$ et leur loi a une densité inconnue θ par rapport à la mesure de Lebesgue. On mettra le plus souvent des restrictions sur θ pour ne pas avoir un espace de paramètres trop gros. Par exemple, on supposera que θ est continue, ou deux fois différentiable, ou croissante, etc. Là encore, il s'agit de modèles dominés. On peut aussi prendre pour paramètre la fonction de répartition des X_i et pour Θ l'ensemble des fonctions de répartition sur $[0;1]$, ou sur $\mathbb{R}$, ou l'ensemble des fonctions de répartition continues sur $[0;1]$. Il s'agit alors d'un modèle qui n'est pas dominé. Dans tous ces cas, l'espace des paramètres n'est pas un sous-ensemble d'un espace euclidien et l'on parlera alors de modèles *non-paramétriques*.

Dans la plupart des exemples vus jusqu'ici, nous avons considéré des variables i.i.d. Ce n'est évidemment pas le seul cas possible, loin de là, mais c'est un type de modèle très répandu que l'on appelle *modèle d'échantillonnage*.

2 Les problèmes statistiques classiques

Nous voulons ici évoquer les problèmes que cherche à résoudre la Statistique en partant du modèle très simple des essais de Bernoulli. Nous supposons donc que les observations

$X_1, \ldots, X_n$ sont i.i.d. sur $\{0;1\}$ avec $\mathbb{P}_\theta(X_i = 1) = \theta$ où $\theta \in \Theta \subset [0;1]$ est un paramètre inconnu. Comme déjà indiqué, on ne dispose généralement que d'un résumé exhaustif des observations, la v.a. binomiale $N = \sum_{i=1}^n X_i$ de loi $\mathcal{B}(n, \theta)$.

2.1 Estimation

Le premier problème que l'on peut se poser est celui d'*estimer* le paramètre θ, c'est à dire de tenter de le deviner à partir des observations. On cherche donc à construire un *estimateur*, c'est à dire une statistique $\widehat{\theta}(X_1, \ldots, X_n)$ ou $\widehat{\theta}(N)$, évidemment indépendante de θ, et telle que $\widehat{\theta}$ soit une *estimation* (c'est à dire une approximation) de la vraie valeur de θ. N'importe quelle variable aléatoire fonction des observations (nimporte quelle statistique) peut faire l'affaire, par exemple l'estimateur déterministe $\widehat{\theta} \equiv 1/2$, mais cela ne présente pas grand intérêt. Ce que l'on souhaite, c'est évidemment que $\widehat{\theta}$ ressemble à θ (penser aux sondages ou aux études médicales) et, plus précisément, que, quelle que soit la vraie valeur (inconnue) de θ, $|\widehat{\theta} - \theta|$ soit petit avec une grande probabilité. Une manière simple d'évaluer les performances d'un estimateur $\widehat{\theta}(\boldsymbol{X})$ d'un paramètre réel θ est de calculer son *risque quadratique*.

Définition 3 *Étant donné une expérience statistique avec $\Theta \subset \mathbb{R}$ et un estimateur $\widehat{\theta}(\boldsymbol{X})$, son risque quadratique est défini par*

$$R(\widehat{\theta}, \theta) = \mathbb{E}_\theta\left[(\widehat{\theta} - \theta)^2\right] = \int_E (\widehat{\theta}(x) - \theta)^2 \, dP_\theta(x).$$

C'est une fonction de Θ dans $[0; +\infty]$.

Rappel (Inégalités de Markov et de Bienaymé - Tchébychev) *Soit Y une v.a. réelle, alors, pour tout $t > 0$, $p > 0$,*

$$\mathbb{P}[|Y| > t] \leq t^{-p}\mathbb{E}[|Y|^p] \qquad et \qquad \mathbb{P}[|Y - \mathbb{E}[Y]| > t] \leq t^{-2}\,\mathrm{Var}(Y).$$

Dans le cas qui nous intéresse, ceci donne pour $p = 2$

$$\mathbb{P}_\theta\left[|\widehat{\theta} - \theta| > t\right] \leq t^{-2}\mathbb{E}_\theta\left[(\widehat{\theta} - \theta)^2\right] = t^{-2}R(\widehat{\theta}, \theta).$$

Donc, si le risque quadratique est petit, $|\widehat{\theta} - \theta|$ est nécessairement petit avec une grande probabilité.

Une décomposition fondamentale du risque quadratique, obtenue en écrivant que $\widehat{\theta} - \theta = (\widehat{\theta} - \mathbb{E}_\theta[\widehat{\theta}]) + (\mathbb{E}_\theta[\widehat{\theta}] - \theta)$, est la suivante :

$$R(\widehat{\theta}, \theta) = \left(\mathbb{E}_\theta[\widehat{\theta}] - \theta\right)^2 + \mathrm{Var}_\theta(\widehat{\theta}). \tag{I.2.1}$$

Définition 4 *Le terme $\mathbb{E}_\theta[\widehat{\theta}] - \theta$ est dit biais de l'estimateur au point θ. Si le biais d'un estimateur est nul, quel que soit $\theta \in \Theta$, l'estimateur est dit* sans biais.

La formule (I.2.1) donne donc la décomposition du risque quadratique en carré du biais et variance de l'estimateur. On pourrait naïvement penser que l'estimateur idéal est celui de biais et variance nuls. Mais un tel estimateur n'existe pas, sauf dans le cas pathologique où toutes les lois P_θ sont deux à deux étrangères. Si l'on suppose, au contraire, que toutes ces lois sont mutuellement absolument continues, un estimateur de variance nulle est constant p.s. (pour toutes les lois) et égal à θ_0 de sorte que son biais est $\theta_0 - \theta$, ce

qui est catastrophique si θ est loin de θ_0. On se contentera de rechercher des estimateurs dont le biais et la variance sont tous deux petits.

Dans le cas du modèle des essais de Bernoulli ou du modèle binomial avec $N = \sum_{i=1}^{n} X_i$, $\mathbb{E}_\theta[N] = n\theta$ et $\mathrm{Var}(N) = n\theta(1-\theta)$, de sorte que l'estimateur $\widehat{\theta}_n = N/n$ a les propriétés suivantes :

i) il est sans biais ;

ii)
$$R(\widehat{\theta}_n, \theta) = \mathrm{Var}_\theta(\widehat{\theta}_n) = n^{-1}\theta(1-\theta) \leq (4n)^{-1} \underset{n\to+\infty}{\longrightarrow} 0;$$

iii)
$$\widehat{\theta}_n \underset{n\to+\infty}{\longrightarrow} \theta \quad \text{en probabilité et } \mathbb{P}_\theta\text{-p.s.};$$

iv) en notant par $\rightsquigarrow$ la convergence en loi,
$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, \theta(1-\theta)) \quad \text{quand } n \to +\infty.$$

Toutes ces propriétés montrent que si n est grand, $\widehat{\theta}_n$ est proche de θ en divers sens.

2.2 Intervalles de confiance

2.2.1 Définitions

Un "bon" estimateur fournit une valeur approximative de la vraie valeur du paramètre mais dans de nombreuses situations on souhaite disposer d'une information plus précise. C'est en particulier vrai dans le cas des sondages électoraux pour lesquels on ne se contentera pas de l'information que le score de Machin sera approximativement de 41%. Ce que l'on recherche généralement c'est une *fourchette électorale*, c'est à dire une prévision de la forme suivante : "Machin aura entre 39 et 43% des voix". Ce sont de telles annonces que l'on entend à la radio ou à la télévision avant les élections ou dès l'arrêt du vote. Ce que les commentateurs omettent de signaler, c'est qu'un tel résultat : $\theta \in [0,39; 0,43]$ n'est pas certain mais doit être précisé et interprété convenablement. Comme on l'a vu, on dispose, dans le modèle binomial, d'un estimateur $\widehat{\theta}_n$ de risque borné par $(4n)^{-1}$, ce qui implique, par l'Inégalité de Markov, que

$$\mathbb{P}_\theta\left[|\widehat{\theta}_n - \theta| > t\right] \leq 1/\left(4nt^2\right) \quad \text{pour tout } \theta \in [0;1],$$

ou, en fixant $t = 1/\left(2\sqrt{n\alpha}\right)$ avec $0 < \alpha < 1$,

$$\mathbb{P}_\theta\left[\widehat{\theta}_n - \frac{1}{2\sqrt{n\alpha}} \leq \theta \leq \widehat{\theta}_n + \frac{1}{2\sqrt{n\alpha}}\right] \geq 1-\alpha \quad \text{pour tout } \theta \in [0;1].$$

On obtient ainsi un encadrement de θ par deux quantités aléatoires observables $\widehat{\theta}_n \pm 1/\left(2\sqrt{n\alpha}\right)$, avec une probabilité garantie $1-\alpha$ que l'on peut régler. Évidemment, plus α est petit, plus l'intervalle est grand : en Statistique, on ne peut gagner sur tous les tableaux.

Définition 5 *Étant donné une expérience statistique d'observation $\boldsymbol{X}$, avec $\Theta \subset \mathbb{R}$ et un nombre $\alpha \in]0;1[$ (généralement petit), on appelle* intervalle de confiance *pour θ de niveau (de confiance) $1-\alpha$ tout intervalle aléatoire $[\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})]$ dont les deux bornes sont des statistiques (donc ne dépendent que de quantités connues) et tel que*

$$\mathbb{P}_\theta\left[\underline{\theta}(\boldsymbol{X}) \leq \theta \leq \overline{\theta}(\boldsymbol{X})\right] \geq 1-\alpha \quad \textit{pour tout } \theta \in \Theta.$$

Plus généralement, quel que soit Θ*, on appellera* domaine de confiance *ou* région de confiance *pour* θ *de niveau* α *tout sous-ensemble aléatoire* $D(\boldsymbol{X})$ *de* Θ*, ne dépendant ni du paramètre ni d'autres quantités inconnues, tel que*

$$\mathbb{P}_\theta\left[\theta \in D(\boldsymbol{X})\right] \geq 1-\alpha \quad \textit{pour tout } \theta \in \Theta.$$

Remarques :

i) $\mathbb{R}$ est un intervalle de confiance trivial mais évidemment sans intérêt. Pour un α donné, un intervalle de confiance est d'autant meilleur qu'il est plus précis.

ii) Il ne faut pas perdre de vue que c'est θ qui est fixé (déterministe) et les bornes de l'intervalle de confiance qui sont variables (aléatoires). On pourrait aussi écrire

$$\mathbb{P}_\theta\left[\underline{\theta}(\boldsymbol{X}) \leq \theta \quad \text{et} \quad \overline{\theta}(\boldsymbol{X}) \geq \theta\right] \geq 1-\alpha \quad \text{pour tout } \theta \in \Theta$$

et, dans le cas général, $\mathbb{P}_\theta\left[D(\boldsymbol{X}) \ni \theta\right] \geq 1-\alpha$ pour tout $\theta \in \Theta$.

iii) La probabilité est relative à l'objet aléatoire $\omega \in \Omega$, ce qui signifie que, quel que soit $\theta \in \Theta$, il existe un sous-ensemble A_θ de Ω avec $\mathbb{P}_\theta(A_\theta) \geq 1-\alpha$ et tel que, si $\omega \in A_\theta$, alors $\underline{\theta}(\boldsymbol{X}(\omega)) \leq \theta \leq \overline{\theta}(\boldsymbol{X}(\omega))$. Lorsque l'on a fait l'expérience, on a recueilli une valeur numérique $\boldsymbol{x} = \boldsymbol{X}(\omega)$ (ou un ensemble de valeurs si $\boldsymbol{X} = (X_1, \ldots, X_n)$) et l'intervalle correspondant $\left[\underline{\theta}(\boldsymbol{x}); \overline{\theta}(\boldsymbol{x})\right]$ est fixé. On ne saurait donc écrire

$$\mathbb{P}_\theta\left[\underline{\theta}(\boldsymbol{x}) \leq \theta \leq \overline{\theta}(\boldsymbol{x})\right] \geq 1-\alpha$$

parce que l'évènement en question : $\underline{\theta}(\boldsymbol{x}) \leq \theta \leq \overline{\theta}(\boldsymbol{x})$ est vrai ou faux, pas aléatoire. On dira simplement que $\left[\underline{\theta}(\boldsymbol{x}); \overline{\theta}(\boldsymbol{x})\right]$ est un intervalle de confiance de niveau $1-\alpha$ pour θ.

iv) Il y a des situations dans lesquelles on ne s'intéresse pas à borner θ des deux côtés, mais où l'on veut seulement avoir une majoration (ou une minoration) pour θ, ce qui correspond à prendre $\underline{\theta}(\boldsymbol{X}) = -\infty$ (respectivement $\overline{\theta}(\boldsymbol{X}) = +\infty$). On parlera alors de *borne supérieure de confiance* (respectivement de *borne inférieure de confiance*) pour θ. Ceci se produit par exemple lorsque l'on cherche à évaluer la toxicité d'un produit : la seule chose qui nous intéresse vraiment est d'avoir une borne supérieure de cette toxicité avec un haut niveau de confiance alors qu'une borne inférieure est sans intérêt. Les intervalles définis par une seule borne de confiance sont dits *intervalles de confiance unilatères* alors que ceux qui sont donnés par deux bornes $\underline{\theta}(\boldsymbol{X})$ et $\overline{\theta}(\boldsymbol{X})$ non triviales sont dits *intervalles de confiance bilatères.*

2.2.2 Comment construire des intervalles de confiance ?

Une manière simple, mais fruste, de construire des intervalles de confiance est d'utiliser, comme on l'a vu, le risque quadratique et l'inégalité de Markov, mais il existe des solutions plus sophistiquées. Dans le cas des variables binomiales, on peut, par exemple, utiliser une inégalité dite de "grandes déviations" due à W. Hoeffding.

Proposition 1 (Inégalité de Hoeffding) *On considère* n *v.a.r. indépendantes, centrées et bornées,* $Z_1, \ldots, Z_n$*, prenant respectivement leurs valeurs dans les intervalles* $[a_i; b_i]$*. Alors,*

$$\mathbb{P}\left[\sum_{i=1}^n Z_i \geq x\right] \leq \exp\left[-\frac{2x^2}{\sum_{i=1}^n (b_i - a_i)^2}\right] \quad \textit{pour tout } x > 0.$$

Par symétrie, en changeant Z_i en $-Z_i$ on obtient aussi :

$$\mathbb{P}\left[\sum_{i=1}^{n} Z_i \leq -x\right] \leq \exp\left[-\frac{2x^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right] \quad \text{pour tout } x>0.$$

Notons ici que les variables doivent être indépendantes, mais pas nécessairement de même loi. Dans le cas i.i.d., $b_i - a_i = l$ est indépendant de i et l'inégalité devient, après un changement de variable $x = z\sqrt{n}$ — comparer ici au Théorème Limite Central et à l'inégalité (I.1.1) —

$$\mathbb{P}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \geq z\right] \leq \exp\left[-\frac{2z^2}{l^2}\right] \quad \text{pour tout } z>0.$$

Ceci s'applique en particulier aux essais de Bernoulli X_i de paramètre θ avec $Z_i = X_i - \theta$ à valeur dans $[-\theta; 1-\theta]$. On obtient alors

$$\mathbb{P}_\theta[N - n\theta \geq x] \leq \exp\left[-2x^2/n\right] \quad \text{ou} \quad \mathbb{P}_\theta[N/n - \theta \geq z] \leq \exp\left[-2nz^2\right]. \tag{I.2.2}$$

En changeant Z_i en $-Z_i$ on a aussi $\mathbb{P}_\theta[\theta - N/n \geq z] \leq \exp\left[-2nz^2\right]$, de sorte que

$$\mathbb{P}_\theta[|N/n - \theta| \geq z] \leq 2\exp\left[-2nz^2\right]. \tag{I.2.3}$$

Il est intéressant de noter que ceci donne un bien meilleur résultat que l'inégalité de Bienaymé-Tchebychev. On en déduit qu'avec une probabilité supérieure à $1-\alpha$, indépendamment de θ,

$$\left|\frac{N}{n} - \theta\right| < \left[\frac{-\log(\alpha/2)}{2n}\right]^{1/2} \quad \text{ou} \quad \frac{N}{n} - \left[\frac{-\log(\alpha/2)}{2n}\right]^{1/2} < \theta < \frac{N}{n} + \left[\frac{-\log(\alpha/2)}{2n}\right]^{1/2}.$$

Une solution alternative, valable seulement si n est assez grand, consiste à utiliser l'approximation gaussienne de la loi binomiale. En effet l'on sait que si Z désigne une variable gaussienne standard et I un intervalle de $\mathbb{R}$,

$$\mathbb{P}_\theta\left[\sqrt{n}\frac{\widehat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}} \in I\right] \underset{n\to+\infty}{\longrightarrow} \mathbb{P}[Z \in I].$$

Il s'ensuit que si a est tel que $\mathbb{P}[|Z| \leq a] = 1-\alpha$,

$$\mathbb{P}_\theta\left[\widehat{\theta}_n - \frac{a\sqrt{\theta(1-\theta)}}{\sqrt{n}} \leq \theta \leq \widehat{\theta}_n + \frac{a\sqrt{\theta(1-\theta)}}{\sqrt{n}}\right] \underset{n\to+\infty}{\longrightarrow} 1-\alpha. \tag{I.2.4}$$

Malheureusement, ceci ne donne pas un intervalle de confiance pour θ car les bornes de l'intervalle obtenu dépendent de θ mais on peut agrandir l'intervalle sans problème puisque l'on sait que $\sqrt{\theta(1-\theta)} \leq 1/2$, ce qui donne finalement

$$\liminf_n\ \mathbb{P}_\theta\left[\widehat{\theta}_n - \frac{a}{2\sqrt{n}} \leq \theta \leq \widehat{\theta}_n + \frac{a}{2\sqrt{n}}\right] \geq 1-\alpha.$$

Donc, pour n assez grand, $\left[\widehat{\theta}_n - a/\left(2\sqrt{n}\right); \widehat{\theta}_n + a/\left(2\sqrt{n}\right)\right]$ est un intervalle de confiance pour θ de niveau approximativement $1-\alpha$.

Remarque : L'approximation gaussienne précédente, bien que très couramment utilisée, ne va pas sans poser quelques problèmes liés au fait que la qualité de l'approximation de la loi binomiale (correctement renormalisée) par la loi gaussienne standard dépend évidemment de n mais aussi de θ, lequel est inconnu. Cette approximation est d'autant moins bonne que θ est proche de 0 ou 1. Donc, si $\widehat{\theta}_n$ est loin de $1/2$, ce qui laisse à penser que θ l'est aussi, mieux vaut n'utiliser cette approximation que pour de grandes valeurs de n.

2.3 Changements de paramètres

Lorsque l'on dispose d'un modèle statistique $\{P_\theta, \theta \in \Theta\}$, on n'a pas nécessairement un paramétrage canonique. Il est clair que si g est une application bijective de Θ sur Λ, on peut aussi bien utiliser le modèle $\{Q_\lambda, \lambda \in \Lambda\}$ avec $\lambda = g(\theta)$ et $Q_\lambda = P_{g^{-1}(\lambda)}$. Un exemple type est le modèle gaussien à variance inconnue, $\{\mathcal{N}(0, \sigma^2), \sigma^2 > 0\}$ que l'on pourrait aussi écrire $\{\mathcal{N}(0, \sigma)\}$ en choisissant l'écart-type σ comme paramètre au lieu de la variance σ^2. Ici, nous noterons toujours $\mathcal{N}(\mu, \sigma^2)$ la loi gaussienne d'espérance μ et **variance** σ^2.

Dans le modèle des essais de Bernoulli, il est de tradition de prendre pour paramètre $\theta = \mathbb{P}[X_i = 1]$, mais on pourrait tout aussi bien paramétrer par $\lambda = \mathbb{P}[X_i = 0] = 1 - \theta$. On pourrait aussi choisir $\lambda = \sqrt{\theta}$. De manière plus générale, on peut aussi s'intéresser à une fonction du paramètre, par exemple $\sqrt{\theta(1-\theta)}$ qui est, au facteur (connu) $\sqrt{n}$ près, l'écart-type de la loi binomiale $\mathcal{B}(n, \theta)$. Dans ce cas l'application g n'est pas bijective mais l'intérêt du problème de l'estimation de $g(\theta)$ subsiste. On sera donc souvent amené à estimer ou trouver des intervalles de confiance non pas pour θ mais pour $\lambda = g(\theta)$.

Si l'application g est continue (cas typique) et si l'on dispose d'un "bon" estimateur $\widehat{\theta}$ de θ, tel que $\widehat{\theta} - \theta$ est petit avec une grande probabilité, il est naturel d'utiliser $\widehat{\lambda} = g(\widehat{\theta})$ comme estimateur de $\lambda = g(\theta)$. Dans un cadre asymptotique, si la suite d'estimateurs $\widehat{\theta}_n$ converge vers θ, la suite correspondante $\widehat{\lambda}_n$ convergera vers λ. Cette substitution est ce que l'on appelle la "méthode delta".

Si l'application g est monotone, disons croissante pour fixer les choses, et si l'on sait construire des intervalles de confiance pour θ, on sait également en construire pour λ. En effet si $\lambda = g(\theta)$ et

$$\mathbb{P}_\theta \left[\underline{\theta}(\boldsymbol{X}) \leq \theta \leq \overline{\theta}(\boldsymbol{X}) \right] \geq 1 - \alpha \quad \text{pour tout } \theta \in \Theta,$$

alors

$$\mathbb{P}_\theta \left[g\left(\underline{\theta}(\boldsymbol{X})\right) \leq \lambda \leq g\left(\overline{\theta}(\boldsymbol{X})\right) \right] \geq 1 - \alpha \quad \text{pour tout } \lambda \in \Lambda.$$

Pour aller plus loin, on a besoin d'un nouveau théorème de probabilités qui permet de résoudre élégamment le problème suivant (entre autres). Soit $X_1, \ldots, X_n$ des v.a.r. i.i.d. d'espérance $\mu \neq 0$ et variance σ^2. On sait qu'alors

$$\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu \qquad \text{et} \qquad \sqrt{n}\left(\overline{X}_n - \mu\right) \rightsquigarrow \mathcal{N}\left(0, \sigma^2\right),$$

donc $\overline{X}_n^{-1} \xrightarrow{P} \mu^{-1}$ par continuité de l'application $x \mapsto x^{-1}$ si $x \neq 0$. Que peut-on alors dire de plus sur $\left(\overline{X}_n^{-1} - \mu^{-1}\right)$? Le Théorème suivant permet de répondre à cette question.

Théorème 1 (Méthode delta) *Soit une suite de variables aléatoires $(Z_n)_{n\geq 1}$ et soit $(a_n)_{n\geq 1}$ une suite de réels positifs tendant vers $+\infty$ et tels que pour un certain $z \in \mathbb{R}$ et une variable aléatoire réelle Z,*

$$a_n(Z_n - z) \rightsquigarrow Z.$$

Soit g une fonction définie sur un voisinage de z et dérivable au point z. Alors

$$a_n[g(Z_n) - g(z)] \rightsquigarrow g'(z)Z.$$

Démonstration : Comme $a_n \to +\infty$, $Z_n \xrightarrow{P} z$ et $g(Z_n) \xrightarrow{P} g(z)$ et, comme g est dérivable au point z, $[g(Z_n) - g(z)]/(Z_n - z) \xrightarrow{P} g'(z)$. On a donc

$$a_n[g(Z_n) - g(z)] = a_n(Z_n - z)[g'(z) + o_P(1)]$$

et l'on conclut en utilisant le Théorème de Slutsky que l'on rappelle ci-après. □

Théorème 2 (Slutsky) *Soit deux suites $(X_n)_{n\geq 1}$ et $(Y_n)_{n\geq 1}$ de v.a. réelles telles que $X_n \rightsquigarrow X$ et $Y_n \xrightarrow{P} c$ où c est une constante. Alors,*

$$X_n + Y_n \rightsquigarrow X + c \qquad et \qquad X_n Y_n \rightsquigarrow cX.$$

Remarque : Si $g'(z) = 0$, $g'(z)Z = 0$ donc la loi limite est la masse de Dirac en 0.

Le Théorème 1 s'applique en particulier aux estimateurs. Si pour tout θ, $a_n(\widehat{\theta}_n - \theta) \rightsquigarrow Z_\theta$ sous la loi P_θ, alors $a_n[g(\widehat{\theta}_n) - g(\theta)] \rightsquigarrow g'(\theta)Z_\theta$. Ceci s'appliquera le plus souvent à des estimateurs asymptotiquement gaussiens. Si $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}\left(0, \sigma^2(\theta)\right)$ sous la loi P_θ, alors $\sqrt{n}[g(\widehat{\theta}_n) - g(\theta)] \rightsquigarrow \mathcal{N}\left(0, g'^2(\theta)\sigma^2(\theta)\right)$.

Dans le problème binomial déjà vu, si $\widehat{\theta}_n = N/n$, on sait que $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, \theta(1-\theta))$. Supposons que l'on veuille estimer la variance asymptotique $\theta(1-\theta)$ de $\sqrt{n}(\widehat{\theta}_n - \theta)$. Il suffit d'estimer $v = \theta(1-\theta)$ par $\widehat{v}_n = \widehat{\theta}_n(1 - \widehat{\theta}_n)$. Comme ici $g(x) = x(1-x)$ et $g'(x) = 1-2x$ on obtiendra que $\sqrt{n}(\widehat{v}_n - v) \rightsquigarrow \mathcal{N}\left(0, (1-2\theta)^2\theta(1-\theta)\right)$ avec la convention que $\mathcal{N}(\mu, 0)$ désigne la mesure de Dirac au point μ qui est la limite en loi de $\mathcal{N}(\mu, \sigma^2)$ quand $\sigma \to 0$.

Application : intervalles de confiance asymptotiques Dans la situation précédente, la variance asymptotique de l'estimateur $\widehat{\theta}_n$ (correctement renormalisé) est $\theta(1-\theta)$ et dépend du paramètre — voir (I.2.4) — de sorte que l'on ne peut l'utiliser pour obtenir des intervalles de confiance asymptotiques. Même si l'on admet que la vraie loi de $\sqrt{n}(\widehat{\theta}_n - \theta)$ est suffisamment proche de la loi limite $\mathcal{N}(0, \theta(1-\theta))$ pour que l'on puisse substituer l'une à l'autre, on ne peut utiliser la loi limite directement pour obtenir des intervalles de confiance pour θ car ce que nous dit l'approximation gaussienne, c'est que (I.2.4) est vraie, mais avec un facteur $\theta(1-\theta)$ inconnu. On peut, comme on l'a vu, le majorer par $1/4$, mais c'est une majoration grossière si θ est loin de $1/2$. Une autre solution, fondée sur le Théorème de Slutsky, est de remarquer que, puisque $\widehat{\theta}_n \xrightarrow{P} \theta$, $\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)} \xrightarrow{P} \sqrt{\theta(1-\theta)}$ par continuité, de sorte que

$$\frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}} = \frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{\sqrt{\theta(1-\theta)}} \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}} \rightsquigarrow \mathcal{N}(0,1).$$

On en déduit, comme pour (I.2.4), que si a est tel que $\mathbb{P}[|Z| \leq a] = 1 - \alpha$,

$$\mathbb{P}_\theta \left[\widehat{\theta}_n - \frac{a\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}}{\sqrt{n}} \leq \theta \leq \widehat{\theta}_n + \frac{a\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}}{\sqrt{n}} \right] \underset{n\to+\infty}{\longrightarrow} 1-\alpha,$$

ce qui fournit un véritable intervalle de confiance pour θ, asymptotiquement de niveau $1-\alpha$. Plus généralement, chaque fois que l'on a une suite d'estimateurs *asymptotiquement gaussienne*, c'est à dire telle que $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}\left(0, \sigma^2(\theta)\right)$, et que la *variance asymptotique* $\sigma^2(\theta)$ est une fonction continue de θ, le même raisonnement (on remplace simplement la fonction $\theta \mapsto \sqrt{\theta(1-\theta)}$ par la fonction $\theta \mapsto \sigma(\theta)$) montre que $\sqrt{n}(\widehat{\theta}_n - \theta)/\sigma(\widehat{\theta}_n) \rightsquigarrow \mathcal{N}(0,1)$, d'où les intervalles de confiance asymptotiques de niveau $1-\alpha$:

$$\left[\widehat{\theta}_n - a\sigma(\widehat{\theta}_n)n^{-1/2}; \widehat{\theta}_n + a\sigma(\widehat{\theta}_n)n^{-1/2}\right].$$

Il faut néanmoins rester conscient du fait qu'un tel résultat peut être vraiment très asymptotique et que le niveau de confiance effectif d'un tel intervalle peut, dans certaines situations, être bien différent de la limite $1-\alpha$, même pour d'assez grandes valeurs de n. Par ailleurs, on constate que pour une valeur de n et un niveau de confiance $1-\alpha$ donnés, la longueur de l'intervalle est proportionnelle à $\sigma(\widehat{\theta}_n)$ et donc d'autant meilleure que la variance asymptotique est plus petite. Lorsque l'on doit comparer plusieurs estimateurs asymptotiquement gaussiens, on peut utiliser ce critère, le meilleur estimateur étant celui qui a la plus petite variance asymptotique puisque c'est lui qui fournira les intervalles de confiance les plus petits, asymptotiquement.

2.4 Tests

2.4.1 La notion de test

La notion de test est très importante en Statistique mais sans doute aussi l'une des plus difficiles à appréhender. Elle est liée au problème suivant : prendre une décision (fondée sur un objet aléatoire $\boldsymbol{X}$) qui ne permet que deux choix possibles, comme "oui" ou "non" et que l'on peut toujours désigner, par commodité, par 0 ou 1. Pour une élection avec seulement deux candidats, il s'agit de déterminer lequel sera élu. Pour un essai médical, on veut savoir si un médicament est suffisamment efficace, ou plus efficace qu'un autre ou si ses effets secondaires sont admissibles, etc.

Comme précédemment, dans chaque cas, on dispose d'une observation $\boldsymbol{X}$ de loi P_θ, $\theta \in \Theta$, et la bonne décision dépend de la vraie valeur (inconnue) du paramètre θ. Pour une élection à deux candidats, si θ est la proportion des électeurs qui votent pour l'un des deux candidats, on peut décider du résultat selon que $\theta \geq 1/2$ ou non. Pour un processus qui doit fonctionner avec une proportion maximale $\overline{\theta}$ d'erreurs ou de défauts, on peut décider selon que la vraie proportion $\theta \leq \overline{\theta}$ ou non et le même raisonnement s'applique aux essais médicaux.

Formellement, le problème se présente de la manière suivante par rapport à l'ensemble Θ des paramètres : on veut choisir entre deux hypothèses, notées respectivement $\boldsymbol{H}_0$ et $\boldsymbol{H}_1$, lesquelles correspondent à des valeurs différentes du paramètre θ : $\boldsymbol{H}_i$ correspond à $\theta \in \Theta_i$ pour $i = 0$ ou 1 avec $\Theta_0 \cap \Theta_1 = \varnothing$ et $\Theta_0 \cup \Theta_1 = \Theta$. Le but du jeu est alors de découvrir si le vrai θ, inconnu, appartient à Θ_0 ou Θ_1 à partir d'une observation $\boldsymbol{X}$ de loi P_θ. Il s'ensuit que le processus de décision sera donné sous la forme d'une fonction de test φ, c'est à dire une application mesurable de $(E, \mathcal{E})$ dans $\{0;1\}$, la décision étant $\varphi(\boldsymbol{X})$.

On décide correctement si et seulement si $\varphi(\boldsymbol{X}) = i$ lorsque $\theta \in \Theta_i$ et incorrectement si $\varphi(\boldsymbol{X}) = 1 - i$. On peut alors distinguer deux types d'erreurs selon que $\theta \in \Theta_0$ ou Θ_1, dont les probabilités sont données par

$$\begin{array}{ll} \mathbb{P}_\theta[\varphi(\boldsymbol{X}) = 1] = \mathbb{E}_\theta[\varphi(\boldsymbol{X})] & \text{si } \theta \in \Theta_0; \\ \mathbb{P}_\theta[\varphi(\boldsymbol{X}) = 0] = \mathbb{E}_\theta[1 - \varphi(\boldsymbol{X})] = 1 - \mathbb{E}_\theta[\varphi(\boldsymbol{X})] & \text{si } \theta \in \Theta_1. \end{array}$$

On voit immédiatement que toutes ces quantités sont déterminées par la fonction $\theta \mapsto \mathbb{E}_\theta[\varphi(\boldsymbol{X})] \in [0;1]$, dite *fonction puissance* (ou *caractéristique opérationnelle*) du test. Idéalement, il faudrait que la fonction puissance soit très petite lorsque $\theta \in \Theta_0$ et très grande (proche de son maximum 1) lorsque $\theta \in \Theta_1$. Malheureusement, ceci est le plus souvent impossible pour la raison suivante : si Θ est connexe, les deux ensembles Θ_0 et Θ_1 ont nécessairement une frontière commune et la propriété demandée impliquerait que la fonction $\theta \mapsto \mathbb{E}_\theta[\varphi(\boldsymbol{X})]$ ait une discontinuité sur cette frontière. Or cette fonction est automatiquement continue dans la plupart des modèles statistiques parce que φ est bornée et l'application $\theta \mapsto P_\theta$ continue (pour une topologie forte sur l'espace des probabilités). Il s'avère donc que, dans une situation typique, il est impossible de rendre l'erreur du test petite, uniformément par rapport à θ.

2.4.2 Le point de vue de Neyman et Pearson

Une approche classique, due à Neyman et Pearson consiste à dissymétriser le problème en privilégiant l'une des deux hypothèses à tester par rapport à l'autre. Le choix de l'hypothèse que l'on veut privilégier dépend évidemment du problème et il est laissé au statisticien. Par convention, on note $\boldsymbol{H}_0$ l'hypothèse que l'on veut privilégier et que l'on désigne comme *l'hypothèse* (à tester).

Règle : *On doit choisir pour $\boldsymbol{H}_0$ celle des deux hypothèses pour laquelle il est le plus grave de se tromper, c'est à dire de la rejeter si elle est vraie.*

On nomme $\boldsymbol{H}_1$ *l'alternative*. Si $\varphi(\boldsymbol{X}) = 0$, on dit que *l'on accepte l'hypothèse*, puisque l'on décide $\boldsymbol{H}_0$ et dans le cas contraire que *l'on rejette l'hypothèse*, puisque l'on choisit $\boldsymbol{H}_1$. Dans ce nouveau schéma, les caractéristiques d'un test donnent lieu à une terminologie qu'il est indispensable de bien connaître.

Définition 6 *Étant donné une hypothèse $\boldsymbol{H}_0$ à tester et une alternative $\boldsymbol{H}_1$ correspondant à une partition de l'ensemble des paramètres Θ en deux ensembles Θ_0 et Θ_1 et une fonction de test φ de E dans $\{0;1\}$, on appellera :*

i) région de rejet du test *l'ensemble $W = \{x \in E \,|\, \varphi(x) = 1\}$. L'ensemble W^c est dit* région d'acceptation *;*

ii) fonction puissance *du test la fonction $\theta \mapsto \mathbb{E}_\theta[\varphi(\boldsymbol{X})]$ de Θ dans $[0;1]$;*

iii) erreur de première espèce *la restriction de la fonction puissance à Θ_0 et* puissance du test *sa restriction à Θ_1*. L'erreur de deuxième espèce *est la restriction à Θ_1 de l'application $\theta \mapsto 1 - \mathbb{E}_\theta[\varphi(\boldsymbol{X})] = \mathbb{P}_\theta[\varphi(\boldsymbol{X}) = 0]$, c'est à dire le complément à 1 de la puissance.*

iv) La taille *du test, $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta[\varphi(\boldsymbol{X}) = 1]$, est le maximum de l'erreur de première espèce.*

v) Étant donné un nombre $\alpha \in [0;1]$, le test est dit de niveau α *si sa taille est majorée par α, c'est à dire si $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta[\varphi(\boldsymbol{X}) = 1] \leq \alpha$.*

En dehors de ces problèmes de terminologie, c'est la philosophie de cette approche qu'il convient de bien appréhender.

But du jeu : On se fixe un niveau $\alpha > 0$, petit, pratiquement un nombre entre 0 et 1/10, le plus souvent simple : 10%, 5%, 2%, 1%, 0,5%, 10^{-3}, ... et l'on cherche un test de niveau α dont la puissance est grande (autant que faire se peut) si l'hypothèse est nettement fausse, c'est à dire pour les θ de Θ_1 dont la distance à Θ_0 n'est pas trop petite (voir les arguments de continuité évoqués ci-dessus).

2.4.3 Retour sur le modèle binomial

Dans les exemples desquels nous sommes partis, les problèmes de tests que nous avons considérés se ramenaient à déterminer, dans le cadre du modèle binomial où l'on observe N de loi $\mathcal{B}(n,\theta)$, si le paramètre θ était plus grand ou plus petit qu'une valeur donnée θ_0 (1/2 dans le cas d'une élection). On peut alors toujours, puisque si $N \sim \mathcal{B}(n,\theta)$, $n-N \sim \mathcal{B}(n,1-\theta)$, se ramener au cas où $\Theta_0 = [0;\theta_0]$ donc $\Theta_1 =]\theta_0;1]$ (sous réserve que Θ soit $[0;1]$ tout entier).

On sait qu'un bon estimateur de θ dans ce cadre est $\widehat{\theta}_n = N/n$, en ce sens qu'il est proche du vrai θ avec une grande probabilité si n est grand. Compte-tenu de cette information, une idée naturelle pour effectuer le test est de rejeter $\boldsymbol{H}_0$ si $\widehat{\theta}_n$ est *trop grand*, c'est à dire si $\widehat{\theta}_n > c_n$ ou $N > nc_n$, donc de fixer $\varphi(N) = \mathbb{1}_{]nc_n,n]}(N)$. Reste à décider où l'on doit placer la frontière c_n pour obtenir un test de niveau α donné.

Attention ! On ne peut pas fixer $c_n = \theta_0$. En effet, si $\theta = \theta_0$, donc l'hypothèse est vraie, et n est grand, on sait que $\widehat{\theta}_n$ se comporte approximativement comme une variable gaussienne d'espérance θ_0, donc que $\mathbb{P}_{\theta_0}\left[\widehat{\theta}_n > \theta_0\right] \simeq 1/2$ quand n est grand, ce qui amène à rejeter l'hypothèse avec une probabilité voisine de 1/2 alors qu'elle est vraie.

Le bon raisonnement est le suivant : la fonction puissance est $\mathbb{P}_\theta[N > nc_n]$ qui est une fonction décroissante de c_n (évident) et croissante de θ (nettement moins évident !). Il s'ensuit que pour obtenir une puissance maximale, il convient de choisir c_n aussi petit que possible tout en respectant la contrainte de niveau qui s'écrit

$$\sup_{\theta\leq\theta_0} \mathbb{P}_\theta[N > nc_n] = \mathbb{P}_{\theta_0}[N > nc_n] \leq \alpha,$$

ce qui donne finalement $c_n = \inf\{x > 0 \,|\, \mathbb{P}_{\theta_0}[N > nx] \leq \alpha\}$. Ce calcul peut se faire par ordinateur (en tenant compte du fait que N ne prend que des valeurs entières) ou, si l'on est prêt à perdre un peu, à partir de l'inégalité de Hoeffding. En effet, par (I.2.2), pour $x > \theta_0$,

$$\mathbb{P}_{\theta_0}[N > nx] = \mathbb{P}_{\theta_0}[N - n\theta_0 \geq n(x-\theta_0)] \leq \exp\left[-2n(x-\theta_0)^2\right],$$

ce qui amène à choisir $c_n = \theta_0 + \sqrt{-\log\alpha/(2n)}$. Pour de grandes valeurs de n, on peut utiliser l'approximation gaussienne de $\mathbb{P}_{\theta_0}[N > nx]$ (en faisant attention qu'on l'utilise pour θ_0 et que n doit être d'autant plus grand que θ_0 est proche de 0 ou 1). On obtient alors, en désignant par Z une variable gaussienne standard et par Φ sa fonction de répartition,

$$\begin{aligned}\mathbb{P}_{\theta_0}[N > nx] &= \mathbb{P}_{\theta_0}\left[\frac{N-n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} > \frac{n(x-\theta_0)}{\sqrt{n\theta_0(1-\theta_0)}}\right] \\ &\simeq \mathbb{P}\left[Z > \frac{\sqrt{n}(x-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}\right] = 1 - \Phi\left(\frac{\sqrt{n}(x-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}\right).\end{aligned}$$

On choisira donc $c_n = \theta_0 + \sqrt{\theta_0(1-\theta_0)/n}\,\Phi^{-1}(1-\alpha)$. Que l'on utilise l'une ou l'autre méthode, on trouve $c_n = \theta_0 + c_\alpha n^{-1/2}$ où $c_\alpha > 0$ désigne une fonction décroissante de α qui dépend de la méthode utilisée.

Si $\theta \in \Theta_1$, la loi des grands nombres implique que $\widehat{\theta}_n \underset{n\to+\infty}{\longrightarrow} \theta > \theta_0$ alors que $c_n \to \theta_0$ quand $n \to +\infty$. Donc quand $n \to +\infty$, $\mathbb{P}_\theta\left[\widehat{\theta}_n > c_n\right] \to 1$ et l'on conclut que, pour tout $\theta \in \Theta_1$, la puissance tend vers 1 lorsque $n \to +\infty$. Attention ! La convergence n'est pas uniforme car la fonction puissance est continue alors que la limite ne l'est pas !

2.4.4 Intervalles de confiance et tests

Il existe une manière canonique de construire des tests à partir d'intervalles de confiance ou, plus généralement, de domaines de confiance.

Proposition 2 *Supposons que pour une expérience statistique on sache construire des domaines de confiance $D(\boldsymbol{X})$ de niveau $1-\alpha$, c'est à dire que*

$$\mathbb{P}_\theta[\theta \in D(\boldsymbol{X})] \geq 1-\alpha \quad \textit{pour tout } \theta \in \Theta.$$

Alors les tests donnés par $\varphi(\boldsymbol{X}) = 1$ si et seulement si $D(\boldsymbol{X}) \cap \Theta_0 = \varnothing$ sont de niveau α.

Démonstration : Soit $\theta \in \Theta_0$, alors

$$\mathbb{P}_\theta[\varphi(\boldsymbol{X}) = 1] = \mathbb{P}_\theta[D(\boldsymbol{X}) \cap \Theta_0 = \varnothing] \leq \mathbb{P}_\theta[\theta \notin D(\boldsymbol{X})] \leq \alpha. \qquad \square$$

Si l'on s'en tient au cas réel, les ensembles Θ_0 usuels vont avoir l'une des trois formes suivantes : $\Theta_0 =]-\infty;\theta_0]$ ou $\Theta_0 = [\theta_0;+\infty[$ ou $\Theta_0 = [\theta_0-\delta;\theta_0+\delta]$ avec $\delta \geq 0$. Dans les deux premiers cas, on parle de *tests unilatères* parce que Θ_1 est connexe et dans le dernier de *tests bilatères* parce que Θ_1 est la réunion de deux intervalles séparés par Θ_0.

Règle : *pour un test unilatère, utiliser des intervalles de confiance unilatères de sens opposé à Θ_0 ; pour un test bilatère, utiliser des intervalles de confiance bilatères.*

Appliquons ceci au modèle binomial dans deux situations : un cas de test unilatère, $\Theta_0 = [0,\theta_0]$ pour lequel $\boldsymbol{H}_0$ correspond à $\theta \leq \theta_0$, et un cas de test bilatère, $\Theta_0 = \{\theta_0\}$ pour lequel $\boldsymbol{H}_0$ correspond à $\theta = \theta_0$. Dans le premier cas, on utilisera des intervalles de confiance de la forme $[\underline{\theta}, 1]$. Si l'on prend $[0, \overline{\theta}]$, on rencontre toujours Θ_0 et le test accepte toujours : c'est sans intérêt. Si l'on prend des intervalles bilatères, on fait moins bien qu'avec des intervalles unilatères. On part de (I.2.2) qui dit que pour un z_α convenable, de la forme $c_\alpha/\sqrt{n}$,

$$\mathbb{P}_\theta[\theta \leq N/n - z_\alpha] \leq \exp\left[-2nz_\alpha^2\right] = \alpha. \tag{I.2.5}$$

Ceci donne un intervalle de confiance pour θ de niveau $1-\alpha$, de la forme $]N/n - z_\alpha, 1]$. On rejette $\boldsymbol{H}_0$ si $\theta_0 \leq N/n - z_\alpha$. On peut évidemment vérifier directement à partir de (I.2.5) que le test a le niveau requis. Dans le second cas, on utilise (I.2.3) qui dit que pour un z'_α convenable (plus grand que z_α),

$$\mathbb{P}_\theta[|\theta - N/n| \geq z'_\alpha] \leq 2\exp\left[-2n(z'_\alpha)^2\right] = \alpha.$$

Ceci donne un intervalle de confiance pour θ de niveau $1-\alpha$, de la forme $]N/n - z'_\alpha, N/n + z'_\alpha[$ et l'on rejette $\boldsymbol{H}_0$ si $|\theta_0 - N/n| \geq z'_\alpha$.

Il est très instructif d'analyser la fonction puissance de ce test, disons dans le premier cas : $\beta(\theta) = \mathbb{P}_\theta[N/n - z_\alpha \geq \theta_0]$. On a vu que $\beta(\theta) \leq \alpha$ si $\theta \leq \theta_0$. Par ailleurs il est bien clair que β est une fonction continue de θ et qu'elle est croissante car $\mathbb{P}[\mathcal{B}(n,\theta) \geq t]$ est croissante en θ. Donc pour θ un petit peu plus grand que θ_0, on a une puissance proche de α, donc loin de 1 (mauvaise). Maintenant, en utilisant la version symétrique de (I.2.5), on sait que $\mathbb{P}_\theta[N/n \leq \theta - z_\alpha] \leq \alpha$, donc $\mathbb{P}_\theta[N/n > \theta - z_\alpha] \geq 1 - \alpha$. Si $\theta \geq \theta_0 + 2z_\alpha$, avec une probabilité au moins $1 - \alpha$, $N/n > \theta_0 + z_\alpha$ et l'on rejette $\boldsymbol{H}_0$. La puissance est donc $\geq 1 - \alpha$ si $\theta \geq \theta_0 + 2z_\alpha = \theta_0 + 2c_\alpha/\sqrt{n}$. On peut conclure que les erreurs de première et de seconde espèce du test sont uniformément bornées par α, sauf pour un intervalle inclus dans $]\theta_0; \theta_0 + 2c_\alpha/\sqrt{n}[$ qui est d'autant plus petit que n est grand. Par ailleurs, on vérifie que plus α est petit, plus z_α est grand, ce qui diminue la région de rejet : $\{N/n \geq \theta_0 + z_\alpha\}$ et donc la puissance du test. Choisir un niveau très petit force davantage l'acceptation de l'hypothèse. A la limite, lorsque le niveau tend vers 0, on ne rejette plus jamais et le test devient sans intérêt.

2.4.5 Niveaux de significations (p-values)

Lorsque vous voulez effectuer un test classique à partir d'un logiciel statistique, vous fournissez au logiciel les données et vous lui indiquez le type de test que vous voulez effectuer. Le logiciel ne vous demande pas à quel niveau vous voulez faire le test et ne vous donne pas une réponse sous la forme "accepte" ou "rejette" mais il vous fournit une valeur α_0 entre 0 et 1 que l'on appelle en anglais *p-value* et en français *niveau de signification*. Vous pouvez vous-même tirer la conclusion : si vous testez à un niveau $\alpha < \alpha_0$, le test accepte l'hypothèse mais il la rejette pour les niveaux $\alpha > \alpha_0$.

Interprétation : *Si α_0 est très petit, disons 10^{-4}, on va toujours rejeter l'hypothèse, sauf à utiliser des niveaux très bas. Si au contraire α_0 est grand, disons $\alpha_0 \geq 1/10$, on acceptera l'hypothèse à tous les niveaux usuels, ce qui signifie qu'elle est raisonnable. L'hypothèse est d'autant plus plausible que le niveau de signification est grand. Dans les cas intermédiaires ($\alpha_0 = 0,03$ par exemple) le résultat est ambigu.*

D'où cela vient-il ? On rappelle que si un test φ a un niveau α, la région de rejet correspondante $W_\alpha = \{x \,|\, \varphi(x) = 1\}$ vérifie par définition $\sup_{\theta\in\Theta_0} P_\theta[W_\alpha] \leq \alpha < \alpha'$ pour tout $\alpha' > \alpha$, de sorte que W_α est aussi une région de rejet acceptable pour tous les tests de niveau $\alpha' > \alpha$. En pratique on utilise des familles croissantes de régions de rejet, c'est-à-dire que $W_{\alpha'} \supset W_\alpha$ si $\alpha' > \alpha$, le cas limite $\alpha = 1$ permettant de choisir $W_1 = E$. Il s'ensuit que pour une observation $\boldsymbol{X}$ donnée, l'ensemble des α pour lesquels $\boldsymbol{X} \in W_\alpha$ est un intervalle qui a une borne inférieure $\alpha_0 = \alpha_0(\boldsymbol{X})$ dite *niveau de signification* ou *p-value* du test. On a donc $\alpha_0(\boldsymbol{X}) = \inf\{\alpha \,|\, \boldsymbol{X} \in W_\alpha\}$. Par définition, si $\alpha < \alpha_0$, $\boldsymbol{X} \notin W_\alpha$ et l'on accepte l'hypothèse et si $\alpha > \alpha_0$, $\boldsymbol{X} \in W_\alpha$ et l'on rejette l'hypothèse. Le plus souvent, ces régions de rejet prennent la forme $W_\alpha = \{x \,|\, T(x) > t_\alpha\}$ (ou $W_\alpha = \{x \,|\, T(x) \geq t_\alpha\}$) où t_α est une fonction décroissante de α, de sorte que $\alpha_0(\boldsymbol{X}) = \inf\{\alpha \,|\, T(\boldsymbol{X}) > t_\alpha\}$ (ou $\inf\{\alpha \,|\, T(\boldsymbol{X}) \geq t_\alpha\}$). Si l'application $\alpha \mapsto t_\alpha$ est inversible (par exemple continue et strictement décroissante), d'inverse t^{-1}, alors $\alpha_0(\boldsymbol{X}) = t^{-1}\big(T(\boldsymbol{X})\big)$.