L3 MAPI3

Simulations stochastiques
Partie Statistiques

L'objectif de cette seconde partie du cours est multiple :

- Renforcer le cours de statistique du deuxième semestre par une étude approfondie de quelques questions statistiques.
- L'accent sera mis sur l'utilisation des thèorèmes limites (pour la thèorie) et sur l'illustration par simulation (pour les TP).
- Nous présenterons aussi une introduction aux tests non-paramétriques, plus délicats du point de vue de la théorie.

Si vous souhaitez approfondir les notions abordées ici, notamment pour un niveau Master, nous vous recommandons la lecture de l'ouvrage suivant.

- Vincent Rivoirard et Gilles Stoltz, *Statistique en action*, Vuibert, seconde édition, 2012.

– 1) Premiers pas en statistiques : le problème du sondage

— 2) Test de Kolmogorov Smirnov

Chapitre I

Premiers pas en statistiques : le problème du sondage

— **Imaginons une situation politique en 2022** : Didier Deschamps se présente aux élections présidentielles en Belgique, et nous sommes le jour du second tour, en fin d'après-midi. Peu de bulletins ont été dépouillés, mais l'institut Iflop a réalisé un sondage à la sortie d'un stade afin d'estimer la proportion des électeurs qui ont voté pour Didier Deschamps.

— Après avoir interrogé 1225 personnes, l'institut de sondage a compté 637 électeurs ayant voté pour Didier Deschamps. Ainsi, 52% des personnes interrogées ont voté pour Didier Deschamps.

— Vous êtes journaliste sur une chaîne de télévision, et vous devez parler des élections. Quel pronostic annoncerez-vous à l'antenne ?

— Notons :
  — $N \approx 10^7$ le nombre total d'électeurs
  — M le nombre total de ceux qui ont voté pour Didier Deschamps parmi les N électeurs.
  — n le nombre de personnes interrogées à la sortie du stade

  On souhaite savoir si la proportion $\theta = M/N$ est plus grande que $1/2$.
— **Mais comment se faire une idée de la valeur de θ (qui est inconnue) ?** Une approche très naturelle consiste à dire que θ doit être "proche" de la proportion $\widehat{\theta}_n$ observée sur l'échantillon des n personnes interrogées.

$$\widehat{\theta}_n \simeq \theta$$

— Plus formellement, en posant $X_i = 1$ si la i-ème personne interrogée a voté pour Didier Deschamps et $X_i = 0$ si elle a voté pour l'autre candidat, la v.a. $\widehat{\theta}_n$ est définie par

$$\widehat{\theta}_n = \frac{\sum_{i=1}^n X_i}{n} .$$

— On dit dans ce cas que la proportion empirique $\widehat{\theta}_n$ est un *estimateur* de la proportion inconnue θ. Le mot *estimateur* signifie que $\widehat{\theta}_n$ est construit à partir de l'échantillon $(X_1, \dots, X_n)$. En particulier, cette quantité est une variable aléatoire. C'est une fonction de l'échnatillon

— **Question** : comment mesurer la proximité de $\widehat{\theta}_n$ et de θ ? Ces deux valeurs sont-elles d'ailleurs *toujours* proches ?

— Non, pas si on choisit un mauvais échantillon de n personnes (par exemple ceux qui se rappellent du 10 juillet 2018). Mais si cet échantillon est choisi "au hasard", alors on pressent que $\widehat{\theta}_n$ sera "souvent" proche de θ. Mais que signifie *souvent* précisément ? Une modélisation mathématique est nécessaire.

— Dans cette partie, nous considérons la modélisation la plus simple : le modèle de Bernoulli. On suppose ainsi que parmi la population de N personnes, **on interroge n personnes "complètement au hasard" et "avec remise".**
Population totale : N, sous population interrogée : n.

— Plus formellement : en numérotant chaque individu de la population par $k \in \{1, \dots, N\}$, le choix d'un échantillon correspond à tirer aléatoirement $K_1, \dots, K_n \overset{i.i.d.}{\sim} \mathcal{U}\big(\{1, \dots, N\}\big)$.

— Pour chaque individu $k \in \{1, \dots, N\}$ de la population, on pose $x_k = 1$ s'il a voté pour DD, et $x_k = 0$ sinon. A l'issue du sondage, on n'observe pas tous les x_k mais seulement les valeurs $x_{K_1}, x_{K_2}, \dots, x_{K_n}$ des n personnes interrogées. On pose $X_i = x_{K_i}$ pour tout $i \in \{1, \dots, n\}$. On pose également $M = \big|\{k \in \{1, \dots, N\} : x_k = 1\}\big| = \sum_{k=1}^N x_k$: le nombre total d'électeurs de DD.

— Dans ce cadre, la proportion d'électeurs dans toute la population qui ont voté pour DD est donnée par $\theta = M/N$, et la proportion d'électeurs dans l'échantillon qui ont voté pour DD est égale à $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Proposition 1. — *Dans ce modèle de sondage avec remise (dit* modèle de Bernoulli*), la suite de variables aléatoires $X_1, \dots, X_n$ est i.i.d., de loi $\mathcal{B}(\theta)$.*
— *Conséquence importante : la variable aléatoire $\sum_{i=1}^n X_i$ suit la loi binomiale $\mathcal{B}(n, \theta)$.*

Attention : on remet à chaque fois la personne choisie dans la population totale sinon le paramètre θ changerait à chaque fois.

— En pratique, on n'interroge jamais deux fois le même individu.
— On peut prendre en compte cette remarque en supposant que le tirage des n personnes dans $\{1, \dots, N\}$ est maintenant **"sans remise".**
— Plus précisément, le sondage consiste à choisir aléatoirement un sous-ensemble $\widehat{S} \subset \{1, \dots, N\}$ uniformément parmi tous les sous-ensembles de $\{1, \dots, N\}$ de cardinal n, au sens suivant : pour tout sous-ensemble $S \subset \{1, \dots, N\}$ de cardinal n,

$$\mathbb{P}_\theta\big(\widehat{S} = S\big) = \frac{1}{\binom{N}{n}} \ .$$

— On estime alors θ par la proportion $\widehat{\theta}_n = \frac{1}{n} \sum_{i \in \widehat{S}} X_i$ d'individus de l'échantillon $\widehat{S}$ qui ont voté pour DD.

Proposition 2. *Considérons le modèle de sondage sans remise ci-dessus, et supposons pour simplifier que $n \leqslant \min\{M, N - M\}$ (hypothèse très raisonable en pratique).*

Alors, la variable aléatoire $\sum_{i \in \widehat{S}} x_i$ suit une loi hypergéométrique : pour tout $k \in \{0, 1, \dots, n\}$,

$$\mathbb{P}\left(\sum_{i \in \widehat{S}} X_i = k\right) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \ .$$

Or, on peut montrer que si M et $N - M$ sont grands et $M/N \approx \theta$, alors la loi hypergéométrique ci-dessus est proche de la loi binomiale $\mathcal{B}(n, \theta)$, au sens où :

Proposition 3 (Convergence de la loi hypergéométrique vers la loi binomiale)**.**
Soit $0 \leqslant k \leqslant n$ deux entiers fixés. Alors,

$$\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \xrightarrow[\substack{M, N-M \to +\infty \\ M/N \to \theta}]{} \binom{n}{k} \theta^k (1 - \theta)^{n-k} \ .$$

Concrètement, cette proposition signifie que lorsque M et $N-M$ sont grands, un tirage sans remise est très proche d'un tirage avec remise puisqu'on retrouve la loi binomiale. La similarité entre tirage avec ou sans remise n'est pas surprenante car si M et $N-M$ sont grands, on s'attend, même avec remise, à tirer avec grande probabilité un échantillon de n personnes qui sont deux à deux distinctes.

— **Dans toute la suite du chapitre**, on suppose qu'on est dans le *modèle de Bernoulli*, c'est-à-dire que les v.a. $X_1, \dots, X_n$ sont i.i.d. de loi $\mathcal{B}(\theta)$.

— En particulier, la somme $\sum_{i=1}^n X_i$ suit la loi binomiale $\mathcal{B}(n, \theta)$, ce qui nous permettra de décrire plus facilement la variable aléatoire

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

— Le modèle de Bernoulli dépend d'un paramètre θ ; on notera $\mathbb{P}_\theta()$, $\mathbf{E}_\theta()$ et $\mathrm{Var}_\theta()$ pour spécifier que ces quantités sont calculées avec le paramètre θ.

Une fois posé le modèle mathématique, le statisticien a pour objectif de retrouver le(s) paramètre(s) du modèle, ou de répondre à des questions qui s'y rattachent.

Dans la suite du chapitre, nous décrirons 3 tâches qu'un statisticien peut être amené à réaliser :

1. Estimation (de la proportion θ).
2. Construction d'un intervalle de confiance (pour la proportion θ).
3. Tests d'hypothèses (du type : est-ce que $\theta \leqslant 1/2$ ou $\theta > 1/2$?).

L'estimateur

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

est très naturel puisqu'il correspond à la proportion d'individus de l'échantillon ayant voté pour DD.

En fait, $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ peut aussi être interprété comme un cas particulier de deux méthodes plus générales :

Proposition 4. *Supposons que les v.a. $X_1, \dots, X_n$ sont i.i.d. de loi $\mathcal{B}(\theta)$ de paramètre θ inconnu. Alors, l'estimateur $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ coïncide avec :*

- *l'estimateur obtenu par la méthode des moments ;*
- *l'estimateur du maximum de vraisemblance.*

— Pour se faire une première idée de la qualité de l'estimation de θ par $\widehat{\theta}_n$, on peut déjà remarquer que, en moyenne, l'estimateur $\widehat{\theta}_n$ est correct :

$$\mathbf{E}_\theta\big(\widehat{\theta}_n\big) = \mathbf{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_\theta(X_i) = \theta \ .$$

— On dit dans ce cas que **l'estimateur est sans biais**

— Ce comportement en moyenne n'est pas forcément rassurant : de grosses erreurs positives pourraient compenser de grosses erreurs négatives.

— En fait, la loi forte des grands nombres montre que, puisque les v.a. X_i sont i.i.d., intégrables et d'espérance θ, l'estimateur $\widehat{\theta}_n$ est asymptotiquement correct avec probabilité 1 :

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\text{p.s.}} \theta \ .$$

— Cela signifie concrètement qu'il existe un événement A tel que $\mathbb{P}(A) = 1$ et :

$$\forall \omega \in A,\ \forall \varepsilon > 0,\ \exists n_0(\omega,\varepsilon),\ \forall n \geqslant n_0(\omega,\varepsilon),\ \big|\widehat{\theta}_n(\omega) - \theta\big| \leqslant \varepsilon \ .$$

— C'est déjà mieux qu'un simple contrôle de l'espérance !

— Subsiste quand même un problème : le nombre $n_0(\omega,\varepsilon)$ d'observations nécessaires pour approcher θ avec une précision ε dépend du résultat $\omega \in \Omega$ de l'expérience aléatoire.

— En pratique, on voudrait savoir ce que vaut n_0 (pour choisir n afin d'assurer une précision ε), mais on ne connait pas ω...

— Une façon d'éviter la dépendance en ω est d'utiliser la loi faible des grands nombres (convergence en probabilité) qui implique que : pour tout $\varepsilon > 0$,

$$\mathbb{P}_\theta\big(\big|\widehat{\theta}_n - \theta\big| > \varepsilon\big) \xrightarrow[n\to+\infty]{} 0 \ .$$

— Par conséquent :

$$\forall \varepsilon > 0,\ \forall \alpha \in (0,1),\ \exists n_0(\varepsilon,\alpha),\ \forall n \geqslant n_0(\varepsilon,\alpha),\ \mathbb{P}_\theta\big(\big|\widehat{\theta}_n - \theta\big| > \varepsilon\big) \leqslant \alpha \ .$$

— Ce résultat est préférable car il indique qu'on peut contrôler la probabilité que $\widehat{\theta}_n$ soit distant de θ de plus de ε dès que le nombre d'observations n est choisi assez grand (sans dépendance en ω).
— Malheureusement, ce résultat est encore très qualitatif car la dépendance de $n_0(\varepsilon, \alpha)$ en ε et α n'est pas quantifiée. Dans la suite, on explique comment sont reliées les quantités n, ε et α.

Remarquons déjà que la variance de $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ décroît quand le nombre n d'observations augmente :

Proposition 5 (Variance de la moyenne empirique $\widehat{\theta}_n$)**.**
Supposons que les v.a. $X_1, \ldots, X_n$ sont i.i.d. de loi $\mathcal{B}(\theta)$ de paramètre θ inconnu. Alors, l'estimateur $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ vérifie :

$$\mathrm{Var}_\theta\big(\widehat{\theta}_n\big) = \frac{\theta(1-\theta)}{n} \ .$$

La proposition suivantepermet de montrer qu'avec grande probabilité, la distance entre $\widehat{\theta}_n$ et θ est au plus de l'ordre de $1/\sqrt{n}$:

Proposition 6 (Contrôle avec grande probabilité de la qualité de $\widehat{\theta}_n$)**.**
Supposons que les v.a. $X_1, \ldots, X_n$ sont i.i.d. de loi $\mathcal{B}(\theta)$ de paramètre θ inconnu. Alors, l'estimateur $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ vérifie : pour tout $\varepsilon > 0$,

$$\mathbb{P}_\theta\big(\big|\widehat{\theta}_n - \theta\big| \leqslant \varepsilon\big) \geqslant 1 - \frac{1}{4n\varepsilon^2} \ . \tag{I.1}$$

En particulier, si on fixe $\alpha \in (0,1)$ et qu'on choisit $\varepsilon = 1/\sqrt{4n\alpha}$, on a :

$$\mathbb{P}_\theta\left(\left|\widehat{\theta}_n - \theta\right| \leqslant \frac{1}{\sqrt{4n\alpha}}\right) \geqslant 1 - \alpha \ . \tag{I.2}$$

Preuve :

Le résultat découle de l'inégalité de Bienaymé-Tchebychev.

Rappels :

Inégalité de Markov : Si X est une v.a. positive ou nulle, alors : $\forall x > 0$, $\mathbb{P}(X \geqslant x) \leqslant \mathbf{E}(X)/x$.

Inégalité de Bienaymé-Tchebychev : Si X est une v.a. de carré intégrable, alors : $\forall x > 0$, $\mathbb{P}\big(\big|X - \mathbf{E}(X)\big| \geqslant x\big) \leqslant \mathrm{Var}(X)/x^2$.

On applique l'inégalité de Bienaymé-Tchebychev à la v.a. $\widehat{\theta}_n$ qui est bien de carré intégrable (car à valeurs dans $[0,1]$) et on obtient :

$$\mathbb{P}_\theta\big(|\widehat{\theta}_n - \theta| > \varepsilon\big) \leqslant \mathbb{P}_\theta\big(|\widehat{\theta}_n - \theta| \geqslant \varepsilon\big) \leqslant \frac{\mathrm{Var}_\theta\big(\widehat{\theta}_n\big)}{\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}$$

car $\mathrm{Var}_\theta\big(\widehat{\theta}_n\big) = \theta(1-\theta)/n \leqslant 1/(4n)$ (étudier la fonction $x \mapsto x(1-x)$ sur $[0,1]$). On en déduit l'inégalité (I.1) en passant au complémentaire. Quant à l'inégalité (I.2), elle s'obtient en remarquant que $1/(4n\varepsilon^2) = \alpha$.

La proposition ci-dessus signifie qu'avec grande probabilité ($\geqslant 1-\alpha$), l'estimateur $\widehat{\theta}_n$ est à distance au plus $1/\sqrt{4n\alpha}$ du paramètre inconnu θ. Cela permet de construire un intervalle aléatoire qui contient la vraie valeur de θ avec probabilité supérieure ou égale à $1-\alpha$ (la preuve est une simple réécriture de (I.2)) :

Proposition 7 (Intervalle de confiance pour la proportion θ)**.**
Supposons que les v.a. $X_1, \ldots, X_n$ sont i.i.d. de loi $\mathcal{B}(\theta)$ de paramètre θ inconnu. Alors, pour tout $\alpha \in (0,1)$, l'intervalle aléatoire

$$\widehat{I}_{n,\alpha} = \left[\widehat{\theta}_n - \frac{1}{\sqrt{4n\alpha}}\,;\widehat{\theta}_n + \frac{1}{\sqrt{4n\alpha}}\right]$$

contient la valeur inconnue θ avec probabilité au moins $1-\alpha$, au sens où

$$\mathbb{P}_\theta\big(\widehat{I}_{n,\alpha} \ni \theta\big) \geqslant 1-\alpha.$$

Remarque 1. *Attention, c'est l'intervalle $\widehat{I}_{n,\alpha}$ qui est aléatoire et non le paramètre θ.*

C'est pourquoi la notation $\mathbb{P}_\theta\big(\widehat{I}_{n,\alpha} \ni \theta\big)$ est peut-être moins ambigue que $\mathbb{P}_\theta\big(\theta \in \widehat{I}_{n,\alpha}\big)$.

Remarque 2 (Influence du nombre n d'observations)**.**
La largeur de l'intervalle de confiance est proportionnelle à $1/\sqrt{n}$, donc plus le nombre n d'observations est grand, et plus notre intervalle de confiance est précis.

C'est logique ! Par ailleurs, le fait qu'on ait $1/\sqrt{n}$ plutôt que $1/n$ implique que pour réduire l'erreur d'un facteur 10 (c-à-d gagner un chiffre significatif), il ne faut pas prendre 10 fois plus, mais 100 fois plus d'observations. Nous n'avons pas le choix, c'est la théorie qui l'impose.

— En passant à l'événement complémentaire, on obtient

$$\mathbb{P}_\theta\big(\widehat{I}_{n,\alpha} \not\ni \theta\big) \leqslant \alpha$$

on dit que l'intervalle de confiance $\widehat{I}_{n,\alpha}$ est *de niveau α*.

— Cela correspond à un niveau de risque, puisqu'il s'agit de la probabilité que l'intervalle de confiance soit incorrect.

— Concrètement, cela signifie que si on répétait l'expérience aléatoire un très grand nombre de fois, alors l'intervalle de confiance obtenu serait incorrect moins de $100 \times \alpha\%$ (environ) des cas.
— En pratique, on n'observe qu'un seul échantillon et on ne calcule qu'un seul intervalle de confiance, mais on espère que l'échantillon observé fait parti des plus de $100 \times (1-\alpha)\%$ échantillons favorables. Ce phénomène est illustré plus tard par une figure.
— Le contrôle de la probabilité $\mathbb{P}_\theta\big(\widehat{I}_{n,\alpha} \not\ni \theta\big) \leqslant \alpha$ est vrai quel que soit le nombre n d'observations. Si le contrôle n'était vrai qu'asymptotiquement (pour $n \to +\infty$), c'est-à-dire, si on avait seulement prouvé l'inégalité

$$\limsup_{n\to+\infty} \mathbb{P}_\theta\big(\widehat{I}_{n,\alpha} \not\ni \theta\big) \leqslant \alpha \ ,$$

alors on dirait que $\widehat{I}_{n,\alpha}$ serait un intervalle de confiance de niveau *asymptotique* α.

— Construire des intervalles de confiance *asymptotiques* peut paraître moins fort que construire des intervalles dont on contrôle le niveau pour tout $n \in \mathbf{N}^*$, mais cela a quand même un intérêt.
— En effet, comme le montre la proposition suivante, l'intervalle de confiance $\widehat{I}_{n,\alpha}$ construit avec l'inégalité de Bienaymé-Tchebychev est trop large quand n est grand.

Proposition 8 (Intervalle de confiance trop large)**.**
Supposons que les v.a. $X_1, \dots, X_n$ sont i.i.d. de loi $\mathcal{B}(\theta)$ de paramètre θ inconnu. Alors, pour tout $\alpha \in (0,1)$,

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\left[\widehat{\theta}_n - \frac{1}{\sqrt{4n\alpha}} \ ; \ \widehat{\theta}_n + \frac{1}{\sqrt{4n\alpha}}\right] \ni \theta\right) \geqslant 1 - \exp\left(-\frac{1}{2\alpha}\right)$$

avec $\exp\big(-1/(2\alpha)\big) \leqslant \alpha$ pour tout $\alpha \in (0,1)$ et $\exp\big(-1/(2\alpha)\big) = o(\alpha)$ quand $\alpha \to 0$.

On a :

$$\begin{aligned}
&\mathbb{P}_\theta\left(\left[\widehat{\theta}_n - \frac{1}{\sqrt{4n\alpha}} \ ; \ \widehat{\theta}_n + \frac{1}{\sqrt{4n\alpha}}\right] \ni \theta\right) \\
&\quad = \mathbb{P}_\theta\left(\left|\sqrt{n}\,\frac{\widehat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}}\right| \leqslant \frac{1}{\sqrt{4\theta(1-\theta)\alpha}}\right) \xrightarrow[n\to+\infty]{} \int_{-\frac{1}{\sqrt{4\theta(1-\theta)\alpha}}}^{\frac{1}{\sqrt{4\theta(1-\theta)\alpha}}} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx \qquad \text{(I.3)}
\end{aligned}$$

d'après le Théorème Central Limite (TCL). En posant $\bar{\Phi}(u) = \displaystyle\int_u^{+\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$, l'intégrale ci-dessus se réécrit $1 - 2\bar{\Phi}(u_\alpha)$ avec $u_\alpha = 1/\sqrt{4\theta(1-\theta)\alpha}$.

Inégalités utiles :

$$\forall u > 0 \ , \qquad \bar{\Phi}(u) \leqslant \frac{e^{-u^2/2}}{u\sqrt{2\pi}} \tag{I.4}$$

$$\forall u \geqslant 1 \ , \qquad \bar{\Phi}(u) \leqslant \frac{e^{-u^2/2}}{2} \ . \tag{I.5}$$

Preuve de la première inégalité : on pose $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$ et on remarque que $\varphi'(x) = -x\varphi(x)$, si bien que

$$\varphi(u) = \int_u^{+\infty} x\varphi(x)dx \geqslant u \int_u^{+\infty} \varphi(x)dx = u\,\bar{\Phi}(u) \ .$$

Ensuite, (I.5) découle [1] de (I.4) en remarquant que $u\sqrt{2\pi} \geqslant \sqrt{2\pi} \geqslant 2$ car $u \geqslant 1$.

Nous sommes maintenant en mesure de majorer l'intégrale appraissant à droite de (I.3). Nous avons dit qu'il s'agissait de $1 - 2\bar{\Phi}(u_\alpha)$. Or, d'après l'inégalité (I.5) ci-dessus et la définition de $u_\alpha = 1/\sqrt{4\theta(1-\theta)\alpha}$,

$$2\bar{\Phi}(u_\alpha) \leqslant e^{-u_\alpha^2/2} = \exp\left(-\frac{1}{8\theta(1-\theta)\alpha}\right) \leqslant \exp\bigl(-1/(2\alpha)\bigr) \ ,$$

où la dernière inégalité découle du fait que la fonction $x \mapsto \exp\bigl(-1/x\bigr)$ est croissante sur $\mathbf{R}_+^*$ et que $8\theta(1-\theta)\alpha \leqslant 2\alpha$ (rappelons que $\theta(1-\theta) \leqslant 1/4$).

— Par ailleurs, pour montrer l'inégalité $\exp\bigl(-1/(2\alpha)\bigr) \leqslant \alpha$, il suffit de poser $x = 1/\alpha$ et de vérifier que $xe^{-x/2} \leqslant 1$ pour tout $x > 0$, ce qui est vrai car $\frac{d}{dx}\bigl(xe^{-x/2}\bigr) = (1 - x/2)e^{-x/2}$, si bien que $x \mapsto xe^{-x/2}$ atteint son maximum en $x = 2$ et ce maximum vaut $2/e \leqslant 1$.

— Enfin, la comparaison $\exp\bigl(-1/(2\alpha)\bigr) \underset{\alpha\to 0}{=} o(\alpha)$ découle du fait que $\lim\limits_{x\to+\infty} xe^{-x/2} = 0$.

— On déduit de la proposition précédente que, quand le nombre n d'observations est grand, l'intervalle de confiance $\widehat{I}_{n,\alpha}$ construit avec l'inégalité de Bienaymé-Tchebychev a un niveau de risque bien plus faible que α .
— Par conséquent, si on est prêt à accepter un niveau de risque asymptotique vraiment égal à α, on peut prendre un intervalle moins large, dicté par le Théorème de la Limite Centrale et le lemme de Slutsky.

1. En fait, l'inégalité (I.5) est vraie pour tout $u \geqslant 0$ mais la preuve est plus délicate.

Proposition 9 (Intervalle de confiance de niveau asymptotique α)**.**
Supposons que les v.a. $X_1, \ldots, X_n$ sont i.i.d. de loi $\mathcal{B}(\theta)$ de paramètre θ inconnu. Alors, pour tout $\alpha \in (0,1)$, si on note $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$ (cf. figure I.1), on a :

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\left[\widehat{\theta}_n - q_{1-\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} \; ; \; \widehat{\theta}_n + q_{1-\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} \right] \ni \theta \right) = 1-\alpha \; .$$

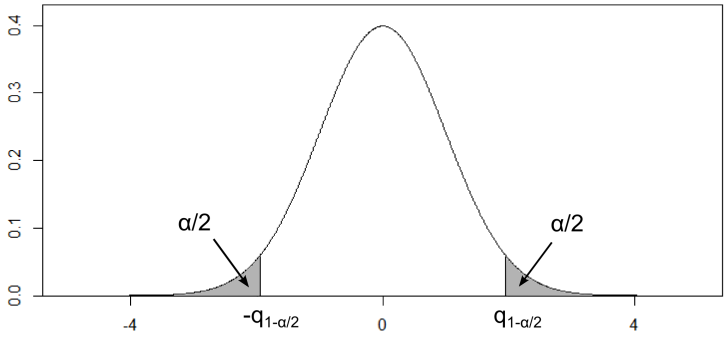

FIGURE I.1 – Représentation du quantile $q_{1-\alpha/2}$ d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$. Par symétrie, toute v.a. $X \sim \mathcal{N}(0,1)$ vérifie $\mathbb{P}(|X| \leqslant q_{1-\alpha/2}) = 1-(\alpha/2+\alpha/2) = 1-\alpha$.

Par le TCL et le lemme de Slutsky (déjà fait dans la première partie du cours). Rappelons à très grands traits la preuve :

$$\begin{aligned}
&\mathbb{P}_\theta\left(\left[\widehat{\theta}_n - q_{1-\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} \; ; \; \widehat{\theta}_n + q_{1-\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} \right] \ni \theta \right) \\
&= \mathbb{P}_\theta\left(\left| \widehat{\theta}_n - \theta \right| \leqslant q_{1-\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} \right) \\
&= \mathbb{P}_\theta\left(\left| \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}} \sqrt{n}\, \frac{\widehat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}} \right| \leqslant q_{1-\alpha/2} \right) \qquad \text{(I.6)}
\end{aligned}$$

Or,

$$\sqrt{n}\,\frac{\widehat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ (TCL)} \quad \text{et} \quad \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}} \xrightarrow[n\to+\infty]{\mathcal{L}} 1 \text{ (car cvgce p.s.)}$$

Par conséquent, d'après le lemme de Slutsky, la probabilité à droite de l'inégalité (I.6) tend vers la probabilité $\displaystyle\int_{-q_{1-\alpha/2}}^{q_{1-\alpha/2}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = 1-\alpha$ (par définition du quantile $q_{1-\alpha/2}$).

— L'intervalle de confiance de la proposition 9 ci-dessus est plus fin que l'intervalle de confiance qu'on avait construit avec l'inégalité de Bienaymé-Tchebychev (car $\widehat{\theta}_n(1-\widehat{\theta}_n) \leqslant 1/4$ et $q_{1-\alpha/2} \leqslant 1/\sqrt{\alpha}$ pour les mêmes raisons que dans la preuve de la proposition 8).
— Ainsi, si on est dans le régime $n \to +\infty$, on a une meilleure précision sur θ. On verra également en TD comment raffiner l'intervalle de confiance tout en préservant un niveau α garanti pour tout $n \in \mathbf{N}^*$ (avec l'inégalité de Hoeffding).

— En pratique, en sciences expérimentales ou en ingénierie, il est fréquent de prendre un niveau $\alpha = 5\% = 0.05$, si bien que puisque $q_{97.5\%} \leqslant 1.96$ (ces valeurs sont tabulées), on en déduit un intervalle de confiance de niveau asymptotique 5% très standard :

$$\left[\widehat{\theta}_n - 1.96\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} \ ; \ \widehat{\theta}_n + 1.96\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}\right]$$

En figure I.2, nous avons simulé 80 échantillons $X_1, \ldots, X_n$ de loi de Bernoulli de paramètre $p = 1/2$ et avons représenté les 80 intervalles de confiance obtenus. Cette figure permet de bien illustrer la notion de *niveau*.

Tests

— Jusqu'à maintenant, nous avons expliqué comment estimer une proportion θ inconnue, et comment fournir des marges d'erreurs $\pm\varepsilon$ pour un niveau de risque α donné.
— Dans ce qui suit, l'objectif ultime n'est plus d'estimer θ, mais de répondre à une question sur θ avec deux réponses au choix.

— Dans l'exemple des élections présidentielles décrit en introduction, la question la plus cruciale est de déterminer, à partir de l'échantillon de sondage, si $\theta \leqslant 1/2$ ou $\theta > 1/2$.
— On dit dans ce cas qu'on cherche à réaliser un test d'hypothèses ; deux hypothèses sont en jeu :

$$\text{"}\theta \in [0, 1/2]\text{"} \quad \text{et} \quad \text{"}\theta \in (1/2, 1]\text{"} \ .$$

— Ces deux hypothèses sont dites *composites* car les ensembles de valeurs possibles ($[0, 1/2]$ ou $(1/2, 1]$) ne sont pas réduits à des singletons.

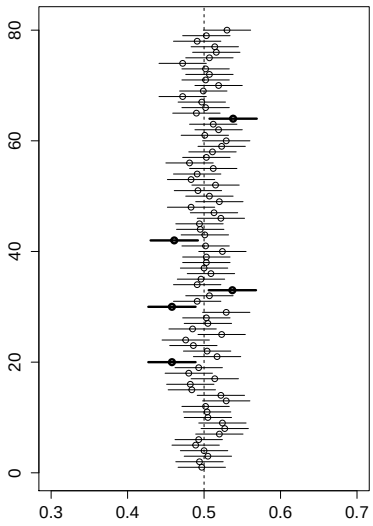

FIGURE I.2 – Visualisation des intervalles de confiance pour θ obtenus avec 80 échantillons (aléatoires) de taille $n = 1000$. Ces intervalles sont construits à partir du Théorème Central Limite et du lemme de Slutsky pour un niveau $\alpha = 0.05$. La vraie valeur de θ est $\theta = 0.5$. Sur les 80 intervalles de confiance obtenus, 5 ne contiennent pas la valeur de θ (en gras). Ainsi, une proportion de $5/80 = 6.25\%$ des échantillons ont donné un mauvais intervalle de confiance. On s'attendait à une proportion de l'ordre de 5% car on a choisi un niveau de risque asymptotique $\alpha = 0.05$.

— Un autre exemple de test de deux hypothèses composites faisant intervenir une proportion θ inconnue peut être tiré du domaine biomédical.
— Ainsi, dans le cadre d'un essai clinique visant à mettre sur le marché (ou non) un nouveau médicament, on peut vouloir tester si la proportion θ d'individus atteints par une certaine pathologie qui seraient guéris par ce nouveau médicament est supérieure au taux de guérison $\theta_0 = 0.7$ observé de longue date pour un médicament qui a déjà fait ses preuves.
— La proportion θ est relative à toute la population de personnes atteintes par cette pathologie. Pour des raisons bioéthiques évidentes, on ne peut pas mettre le médicament tout de suite le marché et attendre d'observer l'effet du nouveau médicament sur toute cette population.
— Un essai clinique est donc réalisé [2] sur un échantillon de taille n ; en d'autres termes, on effectue un "sondage". A partir du taux de guérison $\widehat{\theta}_n$ observé sur l'échantillon, on cherche à identifier laquelle des deux hypothèses composites suivantes est correcte :

2. Le processus d'autorisation de mise sur le marché d'un médicament est un peu plus compliqué ; il y a en fait plusieurs phases d'essais cliniques. Nous avons simplifié volontairement.

$$\text{"}\theta \leqslant \theta_0\text{"} \quad \text{(le nouveau médicament est moins efficace)}$$
$$\text{versus} \quad \text{"}\theta > \theta_0\text{"} \quad \text{(le nouveau médicament est plus efficace)}$$

Comme nous allons le voir tout à l'heure, l'ordre dans lequel on écrit les deux hypothèses n'est pas du tout anodin (les conséquences statistiques sont importantes).

— Remarquons tout d'abord qu'il est vain de vouloir identifier correctement et à coup sûr (avec probabilité 1) laquelle des deux hypothèses
$$\text{"}\theta \leqslant \theta_0\text{"} \quad \text{et} \quad \text{"}\theta > \theta_0\text{"}$$
est vraie.

— En effet, à partir d'un échantillon de taille n, la connaissance qu'on a sur θ n'est pas d'une précision infinie : typiquement, elle se résume à un intervalle de confiance centré en $\widehat{\theta}_n$ et de largeur de l'ordre de $1/\sqrt{n}$.

— Ainsi, si la proportion inconnue θ se situe à distance bien inférieure à $1/\sqrt{n}$ de θ_0, il paraît difficile de pouvoir conclure de façon raisonable entre l'une des deux hypothèses.

— Cet argument très informel peut être réécrit plus rigoureusement (à l'aide de concepts et d'outils de la *théorie de la décision*) pour montrer qu'en effet, cet objectif est trop ambitieux. On reviendra sur cette question en fin de chapitre.

— Commençons par décrire en détail un cadre volontairement plus simple : les deux hypothèses ne sont plus *composites* mais *simples*, c'est-à-dire de la forme
$$H_0 : \text{"}\theta = \theta_0\text{"} \quad \text{et} \quad H_1 : \text{"}\theta = \theta_1\text{"}$$
avec $\theta_0 \neq \theta_1$.

— On se place toujours dans le modèle de Bernoulli, c'est-à-dire que le statisticien dispose d'un échantillon $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{B}(\theta)$ de paramètre θ inconnu.

— Dans ce cadre très simple, on peut introduire sans ambiguité le vocabulaire usuel des tests.

- Un *test de l'hypothèse H_0 contre l'hypothèse H_1* est une fonction mesurable $\varphi_n : \mathcal{X}^n \to \{0,1\}$ qui à un vecteur d'observations $(x_1, \ldots, x_n) \in \mathcal{X}^n$ associe une décision $\varphi_n(x) \in \{0,1\}$. Concrètement, lorsque $\varphi_n(X_1, \ldots, X_n) = 0$, on dira qu'on conserve l'hypothèse H_0, et lorsque $\varphi_n(X_1, \ldots, X_n) = 1$, on dira qu'on rejette l'hypothèse H_0 et qu'on accepte l'hypothèse H_1.

- Le *risque de première espèce* d'un test de deux hypothèses simples est la probabilité de rejeter H_0 alors que H_0 est vraie, ce qu'on peut noter
$$\mathbb{P}_{\theta_0}(\varphi_n(X_1, \ldots, X_n) = 1)$$

- Le *risque de seconde espèce* d'un test de deux hypothèses simples est la probabilité de conserver H_0 alors que H_1 est vraie, ce qu'on peut noter

$$\mathbb{P}_{\theta_1}(\varphi_n(X_1,\dots,X_n)=0)$$

— La *puissance* d'un test de deux hypothèses simples est reliée au risque de seconde espèce : c'est la probabilité d'accepter H_1 quand H_1 est vraie : $\mathbb{P}_{\theta_1}(\varphi_n(X_1,\dots,X_n)=1) = 1-\mathbb{P}_{\theta_1}(\varphi_n(X_1,\dots,X_n)=0)$. La puissance quantifie la capacité du test à détecter l'hypothèse H_1 quand elle est vraie.

— On dit qu'un test de deux hypothèses simples est *de niveau* α ssi son risque de première espèce est inférieur ou égal à α, c-à-d $\mathbb{P}_{\theta_0}(\varphi_n(X_1,\dots,X_n)=1)\leqslant\alpha$.

— Lorsqu'on a seulement un contrôle asymptotique

$$\limsup_{n\to+\infty}\mathbb{P}_{\theta_0}(\varphi_n(X_1,\dots,X_n)=1)\leqslant\alpha\ ,$$

on dit que la suite de tests $(\varphi_n)_{n\in\mathbf{N}^*}$ est *de niveau asymptotique* α.

Résumons par un tableau :

	$\varphi_n=0$	$\varphi_n=1$
$\theta=\theta_0$	OK	erreur de 1ère espèce
$\theta=\theta_1$	erreur de 2nde espèce	OK

Supposons $\theta_0<\theta_1$. Comment à partir de l'observation de

$$X_1,\dots,X_n\sim\mathcal{B}(\theta)$$

peut-on savoir si $H_0:\theta=\theta_0$ ou $H_1:\theta=\theta_1$ est vraie ? Une idée naive serait de comparer $\widehat{\theta}_n$ à $(\theta_0+\theta_1)/2$.

— Si $\widehat{\theta}_n\leqslant\frac{\theta_0+\theta_1}{2}$ on choisit H_0
— Si $\widehat{\theta}_n>\frac{\theta_0+\theta_1}{2}$ on choisit H_1

et donc

$$\varphi_n(X_1,\dots,X_n) = \begin{cases} 0 & \text{si } \widehat{\theta}_n \leqslant \frac{\theta_0+\theta_1}{2} \\ 1 & \text{si } \widehat{\theta}_n > \frac{\theta_0+\theta_1}{2} \end{cases}$$

Problème : imaginons que θ_1 est proche de θ_0 par exemple $\theta_1 = \theta_0 + \frac{c}{\sqrt{n}}$ (pour n suffisamment grand et c petit) on a que le risque de première espèce :

$$\begin{aligned} \mathbb{P}_{\theta_0}(\widehat{\theta}_n > \frac{\theta_0+\theta_1}{2}) &= \mathbb{P}_{\theta_0}(\widehat{\theta}_n > \theta_0 + \frac{c}{2\sqrt{n}}) && \text{(I.7)} \\ &= \mathbb{P}_{\theta_0}\left(\sqrt{n}\frac{\widehat{\theta}_- \theta_0}{\sqrt{\theta_0(1-\theta_0)}} > \frac{c}{2\sqrt{\theta_0(1-\theta_0)}}\right) && \text{(I.8)} \\ &\simeq \int_{\frac{c}{2\sqrt{\theta_0(1-\theta_0)}}}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \quad (TCL) && \text{(I.9)} \end{aligned}$$

Si c est petit la quantité $\frac{c}{2\sqrt{\theta_0(1-\theta_0)}}$ est proche de 0 et

$$\int_{\frac{c}{2\sqrt{\theta_0(1-\theta_0)}}}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \simeq \frac{1}{2}$$

Pour les mêmes raisons le risque de seconde espèce sera proche de 1/2

— Conséquences : si θ_0 et θ_1 sont proches, il faut choisir entre risque de 1ère espèce petit OU risque de 2nde espèce petit.

— Dissymétrie entre H_0 et H_1 : puisqu'il est parfois impossible de contrôler à la fois le risque de 1ère et celui de 2nde espèce, on décide de se concentrer sur l'erreur la plus grave.

Exemples :

- Test d'un nouveau médicament : on cherche à savoir si le taux de guérison θ d'un nouveau médicament est plus grand oùplus petit que celui d'un médicament utilisé depuis plusieurs décennies.

- On suppose que $\theta_0 < \widetilde{\theta} < \theta_1$ correspond à l'efficacité du médicament sachant que $\widetilde{\theta}$ correspond à l'efficacité de l'ancien médicament.

	$\varphi_n = 0$	$\varphi_n = 1$
$\theta = \theta_0$	OK	erreur de 1ère espèce : on choisit le nouveau médicament alors qu'il est moins efficace
$\theta = \theta_1$	erreur de 2nde espèce : on garde l'ancien médicament alors que le nouveau est plus efficace	OK

Dans cet exemple, l'erreur de première espèce est plus grave que celle de seconde espèce.

— Contrôle de qualité : on cherche à savoir si la proportion θ de produits défectueux dans un entrepôt dépasse ou non la norme autorisée par un contrat de vente (0.1% de produits défectueux). Deux rumeurs courent : $\theta = 0.07\%$ et $\theta = 0.15\%$. Si un expert mandaté par l'industriel effectue le test :

$$\begin{aligned} & H_0 : \theta = 0.07 \\ \text{contre} \quad & H_1 : \theta = 0.15 \end{aligned}$$

alors, du point de vue de l'industriel, c'est l'erreur de première espèce qui est la plus grave (car conclure "$\theta = 0.15$" si la réalité est $\theta = 0.07$ le conduirait à revoir pour rien son processus de fabrication, d'oùdes coûts importants inutiles).

— NB : du point de vue du consommateur, c'est l'erreur dans l'autre sens qui est la plus grave. En pratique, on fera toujours en sorte que l'erreur la plus grave corresponde à l'erreur de 1ère espèce. Ainsi, le point du vue du consommateur conduirait à choisir les hypothèses de test suivantes :

$$\begin{aligned} & H_0 : \theta = 0.15 \\ \text{contre} \quad & H_1 : \theta = 0.07 \end{aligned}$$

Dans ce deuxième scénario, c'est bien l'erreur de 1ère espèce qui est la plus grave pour le consommateur, puisqu'elle correspond à ne pas sanctionner l'industriel alors que les produits sont défectueux.

Conclusion : dans tous les cas, on choisira les hypothèses H_0 et H_1 de sorte que **l'erreur la plus grave corresponde à l'erreur de 1ère espèce** (quand c'est possible).

Revenons au cas. de deux hypothèse simples

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta = \theta_1$$

En accord avec le fait que l'on considère l'erreur de première espèce comme la plus grave, on choisit $\alpha \in (0, 1)$ tel que l'on contrôle le risuqe de première espèce par

$$\mathbb{P}_{\theta_0}(\varphi_n(X_1, \ldots, X_n) = 1) \leqslant \alpha$$

ce qui va entraîner une contrainte sur φ_n

Il faut également prendre en compte le fait que l'on veut avoir un risuqe de deuxième espèce petit i.e

$$\mathbb{P}_{\theta_1}(\varphi_n(X_1, \ldots, X_n) = 0) \quad \text{le plus petit possible}$$

En particulier $\varphi_n \equiv 0$ n'est pas un bon choix on a $\mathbb{P}_{\theta_0}(\varphi_n(X_1,\dots,X_n)=1)=0$ mais $\mathbb{P}_{\theta_1}(\varphi_n(X_1,\dots,X_n)=0)=1$

De manière concrète on va choisir judicieusement un nombre t_α tel que

$$\varphi_n = \mathbf{1}_{\{\widehat{\theta}_n > t_\alpha\}}$$

On choisirala règle de décision suivante

$$\begin{array}{lll} "H_0" & \text{si} & \widehat{\theta}_n > t_\alpha \\ "H_1" & \text{si} & \widehat{\theta}_n \leqslant t_\alpha \end{array}$$

Reste à choisir t_α. Posons

$$t_\alpha = \theta_0 + \frac{1}{\sqrt{4n\alpha}}$$

On a d'après B-T

$$\mathbb{P}(\widehat{\theta}_n > t_\alpha) \leqslant \mathbb{P}(|\widehat{\theta}_n - \theta_0| > \frac{1}{\sqrt{4n\alpha}}) \leqslant \alpha$$

On considère donc

$$\varphi_n(X_1,\dots,X_n) = \mathbf{1}_{\widehat{\theta}_n > \theta_0 + \frac{1}{\sqrt{4n\alpha}}}$$

Quid du risque de seconde espèce ?

$$\begin{array}{rcl} \mathbb{P}_{\theta_1}(\widehat{\theta}_n \leqslant t_\alpha) & = & \mathbb{P}_{\theta_1}(\widehat{\theta}_n \leqslant \theta_0 + \frac{1}{\sqrt{4n\alpha}}) \\ & = & \mathbb{P}_{\theta_1}(\widehat{\theta}_n \leqslant \theta_1 - (\theta_1 - \theta_0 - \frac{1}{\sqrt{4n\alpha}})) \\ & \leqslant^{B-T} & \frac{Var_{\theta_1}(\widehat{\theta}_n}{(\theta_1 - \theta_0 - \frac{1}{\sqrt{4n\alpha}})^2)} = \frac{\theta_1(1-\theta_1)}{n(\theta_1 - \theta_0 - \frac{1}{\sqrt{4n\alpha}})^2} \\ & \leqslant & \frac{1}{4n(\theta_1 - \theta_0 - \frac{1}{\sqrt{4n\alpha}})^2} \end{array}$$

Supposons $\theta_1 \geqslant \theta_0 + \frac{1}{\sqrt{4n\alpha}} + \frac{1}{\sqrt{4n\beta}}$ alors

$$\mathbb{P}_{\theta_1}(\widehat{\theta}_n \leqslant t_\alpha) \leqslant \beta$$

Puissance du test : capacité à détecter H_1 si H_1 est vraie. On a sous les conditions ci-dessus

$$\mathbb{P}_{\theta_1}(\widehat{\theta}_n > t_\alpha) = 1 - P_{\theta_1}(\widehat{\theta}_n \leqslant t_\alpha) \geqslant 1 - \beta$$

— Attention, l'interprétation du résultat d'un test est très dangereuse : beaucoup de professionnels/étudiants interprétent mal ces résultats.

— Il faut avoir conscience que puisque les résultats d'un test entraînent souvent des décisions importantes (choix d'un médicament, choix d'une politique, choix d'un investissement économique, etc), l'interprétation est cruciale !

— *La statistique est un outil d'aide à la prise de décisions.*

La difficulté d'interprétation vient du fait qu'un raisonnement logique un peu subtil doit être réalisé. Voici le détail :

- Si les données $x_1, \dots, x_n$ observées conduisent au résultat $\varphi_n(x_1, \dots, x_n) = 1$, il y a de bonnes raisons de croire que H_1 est vraie. En effet : si H_0 était vraie, on aurait $\mathbb{P}_0(\varphi_n(X_1, \dots, X_n) = 1) \leqslant \alpha$, ce qui signifie que la probabilité d'observer un échantillon conduisant à $\varphi_n(X_1, \dots, X_n) = 1$ serait petite. Puisqu'on a observé $\varphi_n(x_1, \dots, x_n) = 1$, par un pseudo-raisonnement par l'absurde, on fait le pari que c'est H_1 qui est vraie. On ne peut pas en être certain (le seul moyen d'être sûr serait de connaître θ, par exemple en faisant un recensement au lieu d'un sondage) ; il s'agit d'un pari au vu des données observées.

- A l'inverse, si les données $x_1, \dots, x_n$ observées conduisent au résultat $\varphi_n(x_1, \dots, x_n) = 0$, la conclusion est beaucoup moins claire. On pourrait en effet être dans l'une des deux situations suivantes :
  — H_0 est vraie et le test a raison ;
  — H_1 est vraie et le test a tort ; malheureusement, la probabilité d'erreur $\mathbb{P}_{\theta_1}(\varphi_n(X_1, \dots, X_n) = 0)$ n'est pas aussi bien contrôlée que celle de l'erreur de première espèce. On ne peut donc pas exclure cette deuxième situation, *sauf si on sait que le risque de deuxième espèce* $\mathbb{P}_{\theta_1}(\varphi_n(X_1, \dots, X_n) = 0)$ *est lui aussi petit, i.e., si la puissance du test est proche*[3] *de* 1.

En conclusion :

- Si le résultat d'un test est $\varphi_n = 1$, on peut conclure que les données observées sont vraisemblablement en désaccord avec l'hypothèse H_0 et donc conclure "vraisemblablement H_1" ;
- Si le résultat d'un test est $\varphi_n = 0$, la conclusion la plus prudente est : "les données ne permettent pas d'exclure l'hypothèse H_0", autrement dit : "peut-être H_0, dans le doute". Mais si on sait par ailleurs que le test est puissant, alors on peut faire le pari que H_0 est vraie.

Les considérations ci-dessus expliquent elles aussi comment choisir les hypothèses H_0 et

3. On verra un peu plus tard dans le cours comment étudier la puissance d'un test. Typiquement, il faut avoir à l'idée qu'un test est d'autant plus puissant que le nombre d'observations est grand.

H_1. En effet, si on veut prouver une hypothèse au travers d'une expérience puis d'un test statistique, il faut mettre cette hypothèse comme H_1 (si on la met comme H_0 et que le résultat du test est $\varphi_n = 0$, ça ne prouve pas forcément grand chose). Le choix des hypothèses est donc le suivant :

H_0 : hypothèse par défaut, de prudence ("hypothèse nulle")

H_1 : hypothèse qui doit à tout prix être démontrée par l'expérience ("hypothèse alternative").

— $H_0 : \theta \leqslant \theta_0 \quad H_1 : \theta > \theta_0$

— Risque de première espèce

$$\sup_{\theta \leqslant \theta_0} \mathbb{P}[\varphi_n(X_1, \ldots, X_n) = 1]$$

— Risque de seconde espèce

$$\sup_{\theta > \theta_0} \mathbb{P}[\varphi_n(X_1, \ldots, X_n) = 0]$$

Proposition 10. *Soit $X_1, \ldots, X_n \sim \mathcal{B}(\theta)$ avec $\theta \in (0,1)$ inconnu et θ_0 fixé alors*

$$\sup_{\theta \leqslant \theta_0} \mathbb{P}_\theta[\widehat{\theta}_n > t] + \sup_{\theta > \theta_0} \mathbb{P}_\theta[\widehat{\theta}_n \leqslant t] = 1,$$

pour tout $t \in \mathbb{R}$.
Le test est donc $\varphi_n = \mathbf{1}_{\widehat{\theta}_n > t}$

— L'application $\theta \to \mathbb{P}_\theta[\widehat{\theta}_n > t]$ est croissante et continue !

— Soit $\theta_0 \leqslant \theta_1$ et soit $U_1, \ldots, U_n \sim \mathcal{U}([0,1])$, on pose $X_i = \mathbf{1}_{U_i \leqslant \theta_0}$ et $Y_i = \mathbf{1}_{U_i \leqslant \theta_1}$ alors on a $X_i \leqslant Y_i$ pour tout i de sorte que

$$\{\frac{1}{n}\sum_{i=1}^n X_i \geqslant t\} \subset \{\frac{1}{n}\sum_{i=1}^n Y_i \geqslant t\}$$

et donc

$$\begin{aligned} \mathbb{P}_{\theta_0}[\widehat{\theta}_n > t] &= \mathbb{P}[\{\frac{1}{n}\sum_{i=1}^n X_i \geqslant t\}] && \text{(I.10)} \\ &\leqslant \mathbb{P}[\{\frac{1}{n}\sum_{i=1}^n Y_i \geqslant t\}] && \text{(I.11)} \\ &= \mathbb{P}_{\theta_1}[\widehat{\theta}_n > t] && \text{(I.12)} \end{aligned}$$

— La continuité en θ se montre en faisant intervenir la loi binomiale car $n\widehat{\theta}_n \sim \mathcal{B}(n, \theta)$.

— Donc on a par continuité et croissance

$$\sup_{\theta \leqslant \theta_0} \mathbb{P}_\theta[\widehat{\theta}_n > t] = \mathbb{P}_{\theta_0}[\widehat{\theta}_n > t] \ (croissance) \tag{I.13}$$

$$\sup_{\theta > \theta_0} \mathbb{P}_\theta[\widehat{\theta}_n \leqslant t] = \mathbb{P}_{\theta_0}[\widehat{\theta}_0 \leqslant t] \ (croissance + continuite) \tag{I.14}$$

et la somme vaut naturellement 1 ce qui conclue la preuve.

— On introduit la fonction puissance d'un test φ_n :

$$\theta \to \mathbb{P}_\theta[\varphi_n = 1]$$

— Puisque $\theta \to \mathbb{P}_\theta[\widehat{\theta}_n > t]$ est croissante le test $\varphi_n = \mathbf{1}_{\theta_n > t_\alpha}$ satisfait
Proposition 11. *Soit $X_1, \ldots, X_n \sim \mathcal{B}(\theta)$ avec $\theta \in (0,1)$ alors le choix de*

$$t_\alpha = \theta_0 + \frac{1}{\sqrt{4n\alpha}}$$

permet d'assurer

$$\sup_{\theta \leqslant \theta_0} \mathbb{P}_\theta[\widehat{\theta}_n > t_\alpha] \leqslant \alpha$$

Par ailleurs

$$\sup_{\theta > \theta_0 \frac{1}{\sqrt{4n\alpha}} + \frac{1}{\sqrt{4n\beta}}} \mathbb{P}_\theta[\widehat{\theta}_n \leqslant t_\alpha] \leqslant \beta$$

— On peut remarquer que pour tout $\theta > \theta_0$

$$\lim_{n \to \infty} \mathbb{P}_\theta[\widehat{\theta}_n > t_\alpha] = 1$$

— On dit dans ce cas que la suite $(\mathbf{1}_{\widehat{\theta}_n > t_\alpha})$ est consistente ou alors on dit que le test est consistent.

— On a

$$\mathbb{P}_\theta[\widehat{\theta}_n > t_\alpha] = 1 - \mathbb{P}_\theta[\widehat{\theta}_n \leqslant t_\alpha] \tag{I.15}$$

$$\geqslant 1 - \frac{1}{\sqrt{4n\left(\theta - \theta_0 - \frac{1}{\sqrt{4n\alpha}}\right)}} \tag{I.16}$$

—

	vertiges	pas de vertige
	108	1055

— Nombre total de patients 1163

— Test $H_0 : \theta \leqslant 8$ *pourcent* $\quad H_1 : \theta > 8$ *pourcent*
— On choisit $\alpha = 0,05$

$$\varphi_n = \mathbf{1}_{\widehat{\theta}_n > 0{,}08 + \frac{1}{\sqrt{4n0{,}05}}}$$

— $\widehat{\theta}_n = 0,0929 < 0,08 + \frac{1}{\sqrt{4n0,05}} = 0,1456$
— Donc on conserve H_0 ou plutôt les observations n'excluent pas que H_0 soit vrai
— On prend un risque α de choisir H_0 vraie.

— On rappelle que l'on pose

$$\Phi(v) = \int_v^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

qui est une fonction décroissante.
— Revenons au test précédent en considérant n grand. Nous avons

$$\mathbb{P}_{\theta_0}\left[\widehat{\theta}_n > \theta_0 + \frac{1}{\sqrt{4n\alpha}}\right] = \mathbb{P}\left[\sqrt{n}\frac{\widehat{\theta}_n - \theta_0}{\sqrt{\theta_0(1-\theta_0)}} > \frac{1}{\sqrt{4\alpha\theta_0(1-\theta_0)}}\right] \quad (I.17)$$

$$\to_n \quad \Phi(\frac{1}{\sqrt{4\alpha\theta_0(1-\theta_0)}}) \quad (I.18)$$

$$\leqslant \quad \Phi(\frac{1}{\sqrt{\alpha}}) \quad car \quad \theta_0(1-\theta_0) \leqslant 1/4 \quad (I.19)$$

$$\leqslant \quad \frac{\exp\left(-\frac{1}{2\alpha}\right)}{2} = o(\alpha) \quad (I.20)$$

— On peut donc choisir t_α plus petit. Avec le quantile $q_{1-\alpha}$ d'ordre α on peut choisir

$$t_\alpha = \theta_0 + q_{1-\alpha}\sqrt{\frac{\theta_0(1-\theta_0)}{n}}$$

avec $\Phi(q_{1-\alpha}) = \alpha$
— Dessin de la densité
— On a alors

$$\sup_{\theta \geqslant \theta_0} \mathbb{P}_\theta[\widehat{\theta}_n > t_\alpha] = \mathbb{P}_{\theta_0}[\widehat{\theta}_n > t_\alpha] \quad (I.21)$$

$$= \mathbb{P}_{\theta_0}\left[\sqrt{n}\frac{\widehat{\theta}_n - \theta_0}{\sqrt{\theta_0(1-\theta_0)}} > \frac{1}{\sqrt{4\alpha\theta_0(1-\theta_0)}}\right] \quad (I.22)$$

$$\to_n \quad \Phi(q_{1-\alpha}) = \alpha \quad (I.23)$$

— En pratique on calcul

$$T = \sqrt{n}\frac{\widehat{\theta}_n - \theta_0}{\sqrt{\theta_0(1-\theta_0)}}$$

où T est parfois appelée variable de test

— Si $T > q_{1-\alpha}$ on rejette H_0 et on fait le pari que H_1 est vrai

— Si $T \leqslant q_{1-\alpha}$ on ne rejette pas H_0

— Retour à l'exemple

vertiges	pas de vertige
108	1055

— Nombre total de patients 1163

— Test $H_0 : \theta \leqslant 8$ *pourcent* $\quad H_1 : \theta > 8$ *pourcent*

— On calcule

$$T = \sqrt{n}\frac{\widehat{\theta}_n - \theta_0}{\sqrt{\theta_0(1-\theta_0)}} \simeq 1,617$$

— Le quantile pour $\alpha = 0,05$ est

$$q_{1-0,05} = 1,645$$

— Ainsi $1,617 \leqslant 1,645$ on ne peut pas rejeter H_0 au vu des observations

— La notion de $p-$valeur est fréquemment utilisée en science expérimentale

— Donnons sa définition dans le cadre d'un test du type

$$\varphi_n = \mathbf{1}_{\widehat{\theta}_n > t_\alpha}$$

Definition 1. *On se place dans le cadre d'un test*

$$H_0 : \theta \leqslant \theta_0 \quad H_1 : \theta > \theta_0$$

et on considère un test du type $\varphi_n = \mathbf{1}_{\widehat{\theta}_n > t_\alpha}$ *où*

$$\alpha \to t_\alpha$$

est une fonction décroissante. On appelle $p-$ *valeur de cette famille de test associée à l'observation* $(X_1, \ldots, X_n)$ *la variable aléatoire*

$$\widehat{\alpha}(X_1, \ldots, X_n) = \sup\{\alpha, \widehat{\theta}_n \geqslant t_\alpha\}$$

— La p-valeur est la première valeur du risque pour lequel on rejette H_0

$$\widehat{\alpha}(X_1,\dots,X_n) = \sup\{\alpha, \widehat{\theta}_n \leqslant t_\alpha\} \quad \text{(I.24)}$$
$$= \inf\{\alpha, \widehat{\theta}_n > t_\alpha\} \quad \text{(I.25)}$$

— Si $\alpha > \widehat{\alpha}$ on a $t_\alpha \leqslant t_{\widehat{\alpha}}$ et donc $\widehat{\theta}_n > t_\alpha$ et on rejette H_0.
— On rejette H_0 quand la $p-$ valeur est petite.
— Attention la $p-$ valeur est une variable aléatoire et donc dépend de notre observation. Pour un autre jeu de donnée cette valeur serait différente.
— **Rq :** Le risque α correspond à la probabilité de rejetter H_0 alors que H_0 est vraie mais il est incorrect d'interpréter la $p-$valeur de la même manière car celle-ci change selon les observations.

— Estimation du paramètre θ dans un modèle de Bernoulli avec un échantillon $X_1,\dots,X_n \sim \mathcal{B}(\theta)$
— Estimateur

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

— C'est un estimateur sans biais

$$\mathbb{E}_\theta[\widehat{\theta}_n] = \theta$$

— La variance est donnée par

$$Var_\theta(\widehat{\theta}_n) = \frac{\theta(1-\theta)}{n}$$

— C'est un estimateur asymptotiquement consistent : LFGN

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\text{p.s.}} \theta \ .$$

— Une fois obtenu une estimation de θ on veut construire un intervalle de confiance de risque α
— Byénaimé Tchebyschev

$$\widehat{I}_{n,\alpha} = \left[\widehat{\theta}_n - \frac{1}{\sqrt{4n\alpha}}\,;\widehat{\theta}_n + \frac{1}{\sqrt{4n\alpha}}\right]$$

— TCL+ Slustky

$$\left[\widehat{\theta}_n - q_{1-\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}\ ;\ \widehat{\theta}_n + q_{1-\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}\ \right]$$

— Test d'hypothèse : $H_0 \quad vs \quad H_1$
— Risque de 1ere espèce : probabilité de rejetter H_0 alors que H_0 est vraie
— Risque de seconde espèce : probabilité de conserver H_0 alors que H_1 est vraie
— Seuil du risque α permet d'obtenir un test du type

$$\varphi_n = \mathbf{1}_{\widehat{\theta}_n > t_\alpha}$$

— Pour déterminer t_α on utilise soit B-T

$$t_\alpha = \theta_0 + \frac{1}{\sqrt{4n\alpha}}$$

— Soit TCL

$$t_\alpha = \theta_0 + q_{1-\alpha}\sqrt{\frac{\theta_0(1-\theta_0)}{n}}$$

Soit $X_1, \ldots, X_n$ des variables aléatoires réelles i.i.d. de densité $f : \mathbf{R} \to \mathbf{R}$ (par rapport à la mesure de Lebesgue sur $\mathbf{R}$) définie par

$$f_\theta(x) = \frac{x^{\theta-1}e^{-x}}{\Gamma(\theta)}\mathbf{1}_{x \geqslant 0}\,,$$

où $\Gamma(\theta) = \int_0^{+\infty} x^{\theta-1}e^{-x}dx$, et où $\theta \in \mathbf{R}_+^*$ est un paramètre inconnu. On note $\mathbf{E}_\theta$, Var_θ et $\mathbb{P}_\theta$ les espérances, variances et probabilités calculées avec cette valeur de θ.

1. Montrer que $\Gamma(t+1) = t \cdot \Gamma(t)$ pour tout $t > 0$.
2. Montrer que $\mathbf{E}_\theta(X_1) = \theta$ et $\mathrm{Var}_\theta(X_1) = \theta$.
3. Montrer que $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur sans biais et fortement consistant du paramètre inconnu θ.
4. Soit $\alpha \in]0,1[$. Construire un intervalle $[A_n, B_n]$ tel que $\mathbb{P}_\theta(A_n \leqslant \theta \leqslant B_n) \xrightarrow[n\to+\infty]{} 1-\alpha$.
5. On suppose ici que $\theta \in]0,9]$. Calculer $\mathrm{Var}_\theta\big(\widehat{\theta}_n\big)$ et en déduire un intervalle de confiance de niveau α pour le paramètre θ, valable pour tout $n \in \mathbf{N}$.

Un de vos amis affirme que moins de la moitié des gens seraient prêts à ramasser votre stylo si vous le laissiez tomber dans un ascenceur. De nature optimiste, vous voulez prouver le contraire.

1. Quelle modélisation du problème parmi les deux suggestions suivantes vous paraît la plus appropriée ? Justifier brièvement, et dites à quoi correspondent θ et θ_0.

$$\begin{cases} H_0 : \theta \leqslant \theta_0 \\ H_1 : \theta > \theta_0 \end{cases} \qquad \text{ou} \qquad \begin{cases} H_0 : \theta > \theta_0 \\ H_1 : \theta \leqslant \theta_0 \end{cases}$$

2. Vous faites l'expérience avec n personnes choisies au hasard (n grand) et calculez la proportion $\widehat{\theta}_n$ de personnes ayant ramassé votre stylo. Donner la formule d'un test φ_n de niveau $\alpha = 5\%$ de H_0 contre H_1 (justifier en 8 lignes maximum).

   Indication : les quantiles d'ordre 90%, 95% *et* 97.5% *de la loi* $\mathcal{N}(0,1)$ *valent respectivement* 1.28, 1.64 *et* 1.96 *environ.*

3. Vous faites l'expérience avec $n = 100$ personnes et observez les résultats suivants :

a ramassé votre stylo	60
n'a pas ramassé votre stylo	40

   Quelle est votre conclusion ?

Chapitre II

Test non-paramétriques

— On va aborder des test dits "non paramétrique" c'est à dire que l'on va tester le jeu d'hypothèse

$$\begin{cases} H_0 : & F = F_{ref} \\ H_1 : & F \neq F_{ref} \end{cases}$$

où F_{ref} correspond à une fonction de répartition (f.d.r) de référence et on a un échantillon

$$X_1, \ldots, X_n \sim F$$

où F est une f.d.r inconnue et les X_i sont i.i.d

— On parle de test non paramétriques car F est décrite par un nombre infini de paramètres $(F(t), t \in \mathbb{R})$

— On rappelle que $F(t) = \mathbb{P}[X_1 \leqslant t] = \mathbb{P}[X_i \leqslant t]$

— L'objet mathématique qui va nous permettre de faire l'estimation de la fonction F inconnue est la fonction de répartition empirique

$$\begin{array}{rccc} F_n : & \mathbb{R} & \to & [0,1] \\ & t & \to & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t} \end{array}$$

— La valeur

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t}$$

dépend des valeurs prises par l'échantillon $X_1, \ldots, X_n$ qui est aléatoire donc la fonction F_n est une fonction aléatoire.

— Si on définit la mesure

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}$$

alors

$$F_n(t) = \mu_n(]-\infty, t]$$

— C'est donc bien une f.d.r, on l'appelle empirique (comme pour la moyenne empirique) car elle dépend des résultats de l'expérience.

— Dessin

— Fixons $t \in \mathbb{R}$ comme les v.a (X_i) sont i.i.d alors les v.a

$$(\mathbf{1}_{X_i \leqslant t})$$

sont également i.i.d et on a que pour tout i

$$\mathbb{E}(\mathbf{1}_{X_i \leqslant t}) = \mathbb{P}[X_i \leqslant t] = \mathbb{P}[X_1 \leqslant t]$$

— On peut donc appliquer la LFGN

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t} \xrightarrow[n\to+\infty]{\text{p.s.}} \mathbb{P}[X_1 \leqslant t] = F(t)$$

— Les v.a $(\mathbf{1}_{X_i \leqslant t})$ sont des v.a i.i.d qui suivent la loi $\mathcal{B}(\mathbb{P}[X_1 \leqslant t]) = \mathcal{B}(F(t))$

— La v.a $nF_n(t)$ compte donc le nombre de succés ainsi

$$nF_n(t) \sim \mathcal{B}(n, F(t))$$

— On a appliqué la LFGN, le TCL nous donne

$$\sqrt{n}\,(F_n(t) - F(t)) \xrightarrow[n\to+\infty]{\text{Loi}} \mathcal{N}(0, F(t)(1-F(t))$$

$$\frac{\sqrt{n}}{\sqrt{F(t)(1-F(t))}}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{1}_{X_i \leqslant t} - F(t))\right) \xrightarrow[n\to+\infty]{\text{Loi}} \mathcal{N}(0,1)$$

— Jusqu'ici on a considéré des résultats lorsque t était fixé et on a obtenu des convergences lorsque F_n était évalué en t.

— Si on fixe un t_0 alors pour un n_0 grand on peut dire des choses sur la proximité de $F_n(t_0)$ et $F(t_0)$ mais si on change et qu'on regarde un temps t_1 alors on aura un autre n_1 qui n'a aucun lien à priori avec le premier n_0.

— Ainsi à chaque $t \in \mathbb{R}$ correspond un n_t suffisamment grand tel que $F_n(t)$ est proche de $F(t)$. Ce résultat n'est pas satisfaisant car on voudrait pouvoir prendre un n suffisament grand indépendant de t. Une sorte d'uniformité.
— On aimerait établir des résultats du type

$$\|Fn - F\| = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to+\infty]{\text{p.s}} 0$$

— On a le résultat suivant appelé parfoirs théorème fondamental de la statistique ou théorème de Glivenko Cantelli
theoreme 1 (Glivenko Cantelli)**.** *Soit* (X_n) *une suite de v.a.i.i.d de f.d.r* F *alors*

$$\|Fn - F\| = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to+\infty]{p.s} 0$$

— On doit donc passer d'une convergence simple à une convergence uniforme

— Maintenant qu'on a un résultat de convergence presque sûre uniforme de F_n vert F on a donc un résultat de consistence. On voudrait maintenant pouvoir faire le test.
— Pour cela nous avons le résultat suivant
theoreme 2 (Kolmogorov)**.** *Soit* (X_n) *une suite de v.a.i.i.d de f.d.r* F *CONTINUE alors la loi de la v.a*

$$D_n = \|F_n - F\|_\infty$$

ne dépend pas de F
— Notons qu'il y a une contrainte c'est la continuité de F.

— Nous donnons ici des éléments de preuve des deux résultats ci-dessus.
— Dans le premier résultat on ne suppose pas F forcèment continue
— Pour établir ces résultats nous allons utiliser l'inverse généralisée de la f.d.r

$$F^{(-1)}(u) = \inf\{t : F(t) \geqslant u\}, u \in]0, 1[$$

— On rappelle que si $U \sim \mathcal{U}([0, 1])$ alors

$$X = F^{(-1)}(U)$$

a pour f.d.r la fonction F.

— Soit $U_1, \dots, U_n$ n v.a.i.i.d $\mathcal{U}([0,1)$ alors

$$X_1, \dots, X_n = F^{(-1)}(U_1), \dots, F^{(-1)}(U_n)$$

sont n v.a.i.i.d de f.d.r la fonction F.

— Ainsi on a

$$\begin{aligned} \|F_n - F\|_\infty = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| &= \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t} - F(t) \right| \\ &\sim \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{F^{(-1)}(U_i)\leqslant t} - F(t) \right| \quad \text{(II.1)} \end{aligned}$$

— Or on a $\{F^{(-1)}(U_i) \leqslant t\} = \{U_i \leqslant F(t)\}$ donc

$$\sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{F^{(-1)}(U_i)\leqslant t} - F(t) \right| = \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant F(t)} - F(t) \right| \quad \text{(II.2)}$$

— Dans le sup ci-dessus on pose $s = F(t)$ on a donc

$$\begin{aligned} \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{F^{(-1)}(U_i)\leqslant t} - F(t) \right| &= \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant F(t)} - F(t) \right| \quad \text{(II.3)} \\ &= \sup_{s\in F(\mathbb{R})} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant s} - s \right| \quad \text{(II.4)} \end{aligned}$$

— En utilisant que le *sup* sur $F(\mathbb{R})$ est plus petit que le sup sur $[0,1]$ on a

$$\begin{aligned} \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{F^{(-1)}(U_i)\leqslant t} - F(t) \right| &= \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant F(t)} - F(t) \right| \quad \text{(II.5)} \\ &= \sup_{s\in F(\mathbb{R})} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant s} - s \right| \quad \text{(II.6)} \\ &\leqslant \sup_{s\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant s} - s \right| \quad \text{(II.7)} \end{aligned}$$

— Grâce à la loi forte des grands nombres on sait

$$\forall s \in [0,1], \exists A_s, t.q, \mathbb{P}(A_s) = 1, \ et \ \forall \omega \in A_s, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant s} \xrightarrow[n\to+\infty]{} s$$

— Pour pouvoir utiliser la phrase

$$\forall s \in [0,1], \exists A_s, t.q, \mathbb{P}(A_s) = 1, \ et \ \forall \omega \in A_s, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant s} \xrightarrow[n\to+\infty]{} s$$

on veut trouver un ensemble A indépendant de s c'est à dire obtenir

$$\exists A, \ t.q \ \forall s \in [0,1], \forall \omega \in A, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega) \leqslant s} \xrightarrow[n\to+\infty]{} s$$

— Problème déjà évoqué l'ensemble A_s dépend de s. Pourquoi n'a t-on pas le droit de considérer

$$\bigcap_{s\in[0,1]} A_s \quad \text{car c'est une intersection non dénombrable}$$

— On va se ramener à un ensemble dénombrable suffisamment gros pour pouvoir trouver cet ensemble A

— On va utiliser le fait que $\mathbb{Q}$ est dénombrable on a donc

$$\forall s \in [0,1] \cap \mathbb{Q}, \exists A_s, t.q, \mathbb{P}(A_s) = 1, \ et \ \forall \omega \in A_s, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega) \leqslant s} \xrightarrow[n\to+\infty]{} s$$

— Ainsi on peut considérer l'ensemble

$$A = \bigcap_{s\in[0,1]\cap\mathbb{Q}} A_s$$

qui est une intersection dénombrable donc c'est un ensemble mesurable.

— Comme on a $\forall s \in [0,1] \cap \mathbb{Q}, \mathbb{P}(A_s) = 1$ on a que $\mathbb{P}(A) = 1$

— Reprenons la phrase

$$\forall s \in [0,1] \cap \mathbb{Q}, \exists A_s, t.q, \mathbb{P}(A_s) = 1, \ et \ \forall \omega \in A_s, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega) \leqslant s} \xrightarrow[n\to+\infty]{} s$$

et regardons ce qui se passe pour notre ensemble $A = \bigcap_{s\in[0,1]\cap\mathbb{Q}} A_s$

— Soit A l'ensemble ci-dessus et soit $s \in [0,1]\cap\mathbb{Q}$ et $\omega \in A$ alors nécessairement$s \in A_s$ et donc

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega) \leqslant s} \xrightarrow[n\to+\infty]{} s$$

— On a donc montré

$$\exists A,\ t.q\ \forall s \in [0,1] \cap \mathbb{Q}, \forall \omega \in A, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} \xrightarrow[n\to+\infty]{} s$$

— Mais on voudrait la même chose sans se restreindre à $\mathbb{Q}$

$$\exists A,\ t.q\ \forall s \in [0,1], \forall \omega \in A, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} \xrightarrow[n\to+\infty]{} s$$

— Pour passer de $s \in [0,1]\cap\mathbb{Q}$ à $s \in [0,1]$, on va utiliser la densité de $\mathbb{Q}$. Soit $s \in [0,1]$, alors il existe deux suites (s_k) et (t_k) telle que

$$s_k \to s,\ et\ t_k \to s$$

et $s_k \leqslant s$ et $t_k \geqslant s$. On peut donc approcher s par dessus et par dessous par des éléments de $\mathbb{Q}$.

— Ainsi $\{U_i \leqslant s_k\} \subset \{U_i \leqslant s\} \subset \{U_i \leqslant t_k\}$ et par suite soit $\omega \in A$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s_k} \leqslant \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} \leqslant \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant t_k}$$

— On considère les limite en n. On doit utiliser la notion de liminf et limsup

$$\begin{aligned} s_k = \lim_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s_k} &\leqslant \liminf_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} \\ &\leqslant \limsup_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} \leqslant \lim_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant t_k} = t_k \end{aligned}$$

— On reprend la dernière inégalité on a donc

$$s_k \leqslant \liminf_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} \leqslant \limsup_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} \leqslant t_k \tag{II.8}$$

— On prend $k \to \infty$ et on a donc par le théorème des gendarmes

$$s = \liminf_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} = \limsup_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} = s$$

— On a donc montré que

$$\lim_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i(\omega)\leqslant s} = s$$

— La conclusion se base sur un théorème dû à Dini.
— Ce résultat dit que si (f_n) est une suite de fonctions croissante i.e $t \to f_n(t)$ est croissante en t et que la suite de fonction converge simplement vers une fonction f quie st continue alors la convergence est uniforme.
— On est exactement dans ce cadre. Soit $\omega \in A$ alors pour tout n la fonction

$$t \to F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{U_i(\omega) \leqslant t}$$

est une fonction croissante en t.
— De plus la suite de fonction (F_n) est une suite de fonction croissante qui converge simplement vers t. Or $t \to t$ est une fonction continue donc le théorème s'applique et
— Attention, on a utilisé à un moment une égalité en loi. Montrer que la convergence presque sûre ci-dessus se transmet malgré l'égalité en loi est assez ardue et nous l'admettrons.

— Revenons au Théorème de Kolmogorov
theoreme 3 (Kolmogorov)**.** *Soit (X_n) une suite de v.a.i.i.d de f.d.r F CONTINUE alors la loi de la v.a*

$$D_n = \|F_n - F\|_\infty$$

ne dépend pas de F
— Nous introduisons

$$G_n(u) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant u}$$

— Nous allons montrer que

$$\|F_n - F\| \sim \|G_n - G\|$$

où $G(u) = u$.
— On reprend

$$\sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{F^{(-1)}(U_i) \leqslant t} - F(t) \right| = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant F(t)} - F(t) \right| \quad \text{(II.9)}$$

$$= \sup_{s \in F(\mathbb{R})} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{U_i \leqslant s} - s \right| \quad \text{(II.10)}$$

$$\text{(II.11)}$$

— On remarque que

$$]0,1[\subset ImF = F(\mathbb{R}) \subset [0,1]$$

car F est continue

— Ainsi on en déduit

$$\sup_{s\in]0,1[}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{U_i\leqslant s}-s\right| \leqslant \sup_{s\in F(\mathbb{R})}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{U_i\leqslant s}-s\right| \leqslant \sup_{s\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{U_i\leqslant s}-s\right|$$

— Or $G_n(1) = G(1)$ et $G_n(0) = G(0)$ donc

$$\sup_{s\in]0,1[}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{U_i\leqslant s}-s\right| = \sup_{s\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{U_i\leqslant s}-s\right|$$

— On a donc

$$\begin{aligned}
\|F_n - F\|_\infty &= \sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{F^{(-1)}(U_i)\leqslant t}-F(t)\right| \\
&= \sup_{s\in F(\mathbb{R})}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{U_i\leqslant s}-s\right| \\
&= \sup_{s\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{U_i\leqslant s}-s\right| \\
&= \|G_n - G\|_\infty \qquad \text{(II.12)}
\end{aligned}$$

— Le résultat en découle

— On a donc besoin de connaître les quantiles de la loi

$$D_n = \|F_n - F\|_\infty$$

— On définit $d_{n,1-\alpha}$ tel que

$$\alpha = \mathbb{P}[D_n > d_{n,1-\alpha}]$$

— Ces valeurs sont connues (des bonnes approximations) tabulées.

— On s'intérresse souvent à D_n pour des valeurs grandes. En particulier il existe W_{max} telle que

$$\sqrt{n}D_n \xrightarrow[n\to+\infty]{\text{Loi}} W_{max}$$

— En particulier pour tout $x \geqslant 0$ et n grand

$$\mathbb{P}[\sqrt{n}D_n \leqslant x] \simeq \mathbb{P}[W_{max} \leqslant x]$$

— On en déduit que

$$\mathbb{P}[D_n > \frac{x}{\sqrt{n}}] \simeq \mathbb{P}[W_{max} > x]$$

et donc

$$d_{n,1-\alpha} \simeq \frac{w_{1-\alpha}}{\sqrt{n}}$$

— Pour n grand les quantiles de W_{max} donnent ceux de D_n

— On est en mesure de présenter le test de Kolmogorov qui va nous permettre de tester

$$\begin{cases} H_0 : & F = F_{ref} \\ H_1 : & F \neq F_{ref} \end{cases}$$

— On pose le test

$$\varphi_n(X_1, \ldots, X_n) = \mathbf{1}_{D_n > d_{n,1-\alpha}},$$

avec $D_n = \|F_n - F_{ref}\|_\infty$

— Le résultat reste vrai même si F_{ref} n'est pas continue.

— On montrera en T.D que ce test est consistent i.e sous H_1

$$\mathbb{P}[\varphi_n(X_1, \ldots, X_n) = 1] \xrightarrow[n\to+\infty]{} 1$$

Soit $(X_1, \ldots, X_n)$ un n-échantillon de fonction de répartition F inconnue. Notre seule connaissance sur F est qu'elle est continue. On note $\mathbf{F}_n$ la fonction de répartition empirique associée à l'échantillon :

$$\forall t \in \mathbf{R}, \qquad \mathbf{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t}\,.$$

On notera F^{-1} l'inverse généralisée de la fonction F.

1. On rappelle que si $U_1, \ldots, U_n$ sont des variables aléatoires i.i.d. uniformes sur $[0, 1]$, alors $\big(F^{-1}(U_1), \ldots, F^{-1}(U_n)\big)$ est un n-échantillon de fonction de répartition F. En déduire que la loi de $\|\mathbf{F}_n - F\|_\infty$ est la même quelle que soit F continue.

   Dans la suite, pour tout $\alpha \in]0, 1[$, on notera $d_{n,1-\alpha}$ un quantile d'ordre $1 - \alpha$ de cette loi.

2. En cours, nous avions construit un intervalle de confiance centré en la moyenne empirique qui contenait le vrai paramètre θ d'une loi de Bernoulli avec probabilité supérieure ou égale à $1 - \alpha$. En vous inspirant de ce résultat, construisez une région de confiance $\widehat{\mathcal{F}}_n \subset [0, 1]^{\mathbf{R}}$ qui contienne la vraie fonction de répartition F avec probabilité supérieure ou égale à $1 - \alpha$.

1. Application : vous observez les valeurs d'échantillon suivantes :

$$0.38 \quad 0.16 \quad 0.04 \quad 0.63 \quad 0.44 \quad 0.27 \quad 0.51 \quad 0.32 \quad 0.06 \quad 0.13$$

2. Représentez graphiquement la fonction $t \mapsto \mathbf{F}_n(t;\omega)$ sur cet exemple. Expliquez brièvement comment on pourrait représenter sur le graphique la région de confiance $\widehat{\mathcal{F}}_n(\omega)$, pour le niveau de risque $\alpha = 5\%$.
   On donne quelques valeurs approchées du quantile $d_{10,1-\alpha}$:

α	0.01	0.025	0.05	0.1
$d_{10,1-\alpha}$	0.489	0.445	0.409	0.369

3. Que signifie très concrètement la valeur 5% ?