Introduction

Dans le jeu Pokémon, le joueur peut capturer les pokémons peuplant l'univers. Ces pokémons possèdent plusieurs caractéristiques, notamment les *statistiques* et la *nature*, qui ont une influence sur elles. Un exemple est donné en annexe. Tout pokémon possède 6 statistiques :

— L'Attaque
— La Défense
— L'Attaque Spéciale
— La Défense Spéciale
— La Vitesse
— Les Points de Vie (PV)

La nature d'un pokémon donne un bonus de $+10\%$ de valeur de statistique à l'une de ses statistiques parmi les 5 premières (les PV ne sont pas affectés par la nature) en contrepartie d'un malus de -10% de valeur de statistique à une autre de ces statistiques (ou bien dans la même statistique, dans ce cas, la nature a un effet neutre, c'est-à-dire qu'elle ne modifie aucune valeur de statistique). Le pokémon peut alors avoir 25 natures, que l'on peut résumer dans ce tableau :

Malus / *Bonus*	Attaque	Défense	Attaque Spéciale	Défense Spéciale	Vitesse
Attaque	*Hardi*	Solo	Rigide	Mauvais	Brave
Défense	Assuré	*Docile*	Malin	Lâche	Relax
Attaque Spéciale	Modeste	Doux	*Pudique*	Foufou	Discret
Défense Spéciale	Calme	Gentil	Prudent	*Bizarre*	Malpoli
Vitesse	Timide	Pressé	Jovial	Naïf	*Sérieux*

Dans cette étude, nous nous intéresserons donc à la répartition des différentes natures chez les pokémons, afin de vérifier si ces natures sont distribuées selon une loi uniforme.

1 Modélisation statistique, contexte de recueil des données

Cette étude porte sur 900 pokémons capturés dans la version *Violet* du jeu. On posera alors $n = 900$. La nature du pokémon $i \in [\![1; n]\!]$ obtenue lors de la capture sera modélisée par une variable aléatoire

$$\tilde{X}_i \quad : \quad (\Omega, \mathscr{A}, \mathbb{P}) \quad \longrightarrow \quad (N, \mathscr{P}(N))$$

où $(\Omega, \mathscr{A}, \mathbb{P})$ désigne l'espace probabilisé du jeu pokémon et N désigne l'ensemble des natures : $N = \{\text{Hardi, Solo, }\ldots\text{,Sérieux}\}$, avec $|N| = 25$. Pour pouvoir considérer des variables aléatoires réelles, il est adéquat de poser une numérotation de notre ensemble N. Pour cela, introduisons la numérotation σ définie comme la numérotation de gauche à droite et de haut en bas dans le tableau donné plus haut :

$$\begin{array}{rccc} \sigma \ : & N & \longrightarrow & [\![1; 25]\!] \\ & \text{Hardi} & \longmapsto & 1 \\ & \text{Solo} & \longmapsto & 2 \\ & & \vdots & \\ & \text{Naïf} & \longmapsto & 24 \\ & \text{Sérieux} & \longmapsto & 25 \end{array}$$

Nous avons alors une nouvelle collection de variables aléatoires :

$$X_i = \sigma(\tilde{X}_i) \quad : \quad (\Omega, \mathscr{A}, \mathbb{P}) \quad \longrightarrow \quad (\llbracket 1;25 \rrbracket, \mathscr{P}(\llbracket 1;25 \rrbracket))$$

qui, cette fois-ci, sont réelles. Évidemment, la numérotation σ n'a rien de « canonique » mais nous verrons que les estimateurs que nous construirons ne dépendront pas de cette numérotation.

J'ai capturé ces pokémons avec, en pokémon de tête de mon équipe, un pokémon ne possédant pas le talent *Synchro*, qui donne automatiquement au pokémon capturé la même nature que le pokémon de tête. Afin de ne pas perdre trop de temps à récolter les données, j'ai capturé les pokémons sans me préoccuper de leur espèce, en pensant que la répartition des natures était la même pour tous les pokémons. Ainsi, pour cette étude, les variables aléatoires $(X_i)_{i\in\llbracket 1;n \rrbracket}$ seront supposées indépendantes et identiquement distribuées, de sorte à ce que, pour la variable aléatoire $X = (X_1, \dots, X_n)$, on puisse poser le modèle statistique suivant :

$$\left(\llbracket 1;25 \rrbracket^n, \mathscr{P}(\llbracket 1;25 \rrbracket^n), ((p_\theta)^{\otimes n})_{\theta\in\Theta}\right)$$

où $\Theta =]0,1[^{25}$ et, pour $\theta \in \Theta$, p_θ désigne la loi discrète sur $(\llbracket 1;25 \rrbracket, \mathscr{P}(\llbracket 1;25 \rrbracket))$ telle que :

$$p_\theta(\{k\}) = \theta_k, \quad \forall k \in \llbracket 1;25 \rrbracket.$$

2 Estimateurs ponctuels des paramètres

Les paramètres d'intérêt de notre loi inconnue sont les θ_k pour $k \in \llbracket 1;25 \rrbracket$. Pour les estimer, on peut utiliser une méthode *plug-in* ou déterminer les estimateurs du maximum de vraisemblance de notre loi. Nous vérifierons qu'en réalité ces deux méthodes donnent les mêmes résultats.

2.1 Estimateurs *plug-in* des paramètres

Pour estimer les paramètres θ_k, nous pouvons utiliser tout simplement la définition de nos paramètres θ_k :

$$p_\theta(\{k\}) = \theta_k \quad \forall k \in \llbracket 1;25 \rrbracket.$$

Ainsi, en notant $\hat{\theta_{kn}}$ notre estimateur *plug-in* de θ_k, on a :

$$\boxed{\hat{\theta_{kn}} = \hat{p}_n(\{k\}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(\{k\}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i=k}.}$$

Il s'agit en réalité des proportions empiriques dans notre échantillon, ainsi, ils ne dépendent pas de la numérotation choisie !

2.2 Estimateurs du maximum de vraisemblance

Notre modèle statistique est un modèle discret, donc dominé par la mesure de comptage sur $\llbracket 1;25 \rrbracket^n$. De plus, étant donné que notre modèle est d'échantillonnage, la vraisemblance s'écrit :

$$L(x_1, \dots, x_n, \theta) = \prod_{i=1}^{n} \theta_{x_i} = \prod_{k=1}^{25} \theta_k^{\sum_{i=1}^{n} \mathbb{1}_{x_i=k}}.$$

En notant ℓ la log-vraisemblance du modèle, on a :

$$\ell(x,\theta) = \sum_{k=1}^{25} \left(\ln(\theta_k) \sum_{i=1}^{n} \mathbb{1}_{x_i=k} \right).$$

Du fait de la stricte concavité de la fonction ln, la log-vraisemblance du modèle est également strictement concave. Cependant, il faut faire attention, car le maximum en θ de notre fonction ℓ doit être trouvé sur l'ensemble $A = \left\{ \theta \in \Theta \;\middle|\; \sum_{k=1}^{25} \theta_k = 1 \right\} = F^{-1}(\{0\})$, où F est définie ainsi :

$$\begin{array}{rccc} F & : & \Theta & \longrightarrow & \mathbb{R} \\ & & \theta & \longmapsto & \displaystyle\sum_{k=1}^{25} \theta_k - 1. \end{array}$$

Nous pouvons alors trouver le maximum en θ de ℓ sur A grâce au théorème des extrema liés : si ℓ admet un maximum θ^* sur A, alors on doit avoir :

$$\exists \lambda \in \mathbb{R}, \quad \nabla_\theta \ell(x,\theta^*) = \lambda \nabla F(\theta^*),$$

c'est-à-dire :

$$\exists \lambda \in \mathbb{R}, \quad \forall k \in [\![1;25]\!], \quad \frac{1}{\theta_k^*} \sum_{i=1}^{n} \mathbb{1}_{x_i=k} = \lambda.$$

Ainsi, on obtient $\theta_k^* = \dfrac{1}{\lambda} \displaystyle\sum_{i=1}^{n} \mathbb{1}_{x_i=k}$, et la condition $\theta^* \in A$ donne :

$$1 = \sum_{k=1}^{25} \frac{1}{\lambda} \sum_{i=1}^{n} \mathbb{1}_{x_i=k} = \frac{1}{\lambda} \sum_{i=1}^{n} 1 = \frac{n}{\lambda}.$$

D'où :

$$\forall k \in [\![1;25]\!], \quad \theta_k^* = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i=k}.$$

Il s'agit bien d'un maximum étant donné que la fonction ℓ est concave sur A qui est un ensemble convexe (car $1 = t + (1-t)$ pour tout $t \in [0,1]$ et Θ est convexe). Ainsi, en notant $\hat{\theta}_{k_n}^{\text{EMV}}$ l'estimateur du maximum de vraisemblance de θ_k, on retrouve :

$$\boxed{\hat{\theta}_{k_n}^{\text{EMV}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i=k} = \hat{\theta}_{kn}.}$$

En utilisant ces formules et en notant $\omega \in \Omega$ le cadre de notre expérience, on obtient alors :

k	Effectif	$\hat{\theta_{kn}}(\omega)$
1 (Hardi)	36	$0,04$
2 (Solo)	39	$0,04333$
3 (Rigide)	35	$0,03889$
4 (Mauvais)	32	$0,03556$
5 (Brave)	38	$0,04222$
6 (Assuré)	38	$0,04222$
7 (Docile)	40	$0,04444$
8 (Malin)	36	$0,04$
9 (Lâche)	32	$0,03556$
10 (Relax)	34	$0,03778$
11 (Modeste)	37	$0,04111$
12 (Doux)	29	$0,03222$
13 (Pudique)	39	$0,04333$
14 (Foufou)	38	$0,04222$
15 (Discret)	35	$0,03889$
16 (Calme)	41	$0,04556$
17 (Gentil)	30	$0,03333$
18 (Prudent)	42	$0,04667$
19 (Bizarre)	36	$0,04$
20 (Malpoli)	47	$0,05222$
21 (Timide)	31	$0,03444$
22 (Pressé)	29	$0,03222$
23 (Jovial)	32	$0,03556$
24 (Naïf)	30	$0,03333$
25 (Sérieux)	44	$0,04889$

Les résultats ont été arrondis à 10^{-5} près.

2.3 Propriétés des estimateurs construits

Comme remarqué précédemment, les estimateurs construits correspondent aux proportions empiriques de nos différentes natures. Nous pouvons alors déduire de cette observations plusieurs propriétés concernant ces estimateurs :

Proposition (Consistance et lois limites)**.** *Pour tout $k \in [\![1;25]\!]$, les variables aléatoires $(\mathbb{1}_{X_i=k})_{1\leqslant i\leqslant n}$ sont indépendantes et suivent la loi $\mathscr{B}(\theta_k)$. Ainsi, pour tout $k \in [\![1;25]\!]$:*

1. *$\hat{\theta_{kn}}$ est un estimateur fortement consistant de θ_k.*
2. *$\hat{\theta_{kn}}$ est asymptotiquement normal, de vitesse $\sqrt{n}$ et de variance $\theta_k(1-\theta_k)$.*

Démonstration. Rappelons que, pour $k \in [\![1;25]\!]$ notre estimateur $\hat{\theta_{kn}}$ est défini ainsi :

$$\hat{\theta_{kn}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i=k}$$

1. Pour tout $\theta \in \Theta$ et pour tout $k \in [\![1;25]\!]$, les variables aléatoires $(\mathbb{1}_{X_i=k})_{1\leqslant i\leqslant n}$ sont indépendantes, identiquement distribuées et suivent, lorsque X_1 suit la loi p_θ, la loi $\mathscr{B}(\theta_k)$. Elles sont donc

d'espérance finie (et également de variance finie), et la loi forte des grands nombres s'applique. Ainsi :

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \hat{\theta}_{kn} \xrightarrow[n\to+\infty]{\mathbb{P}_\theta\text{-p.s.}} \mathbb{E}_\theta(\mathbb{1}_{X_1=k}) = \mathbb{P}_\theta(X_1 = k) = \theta_k.$$

2. Le fait également que les variables $\mathbb{1}_{X_i=k}$ soient de variance finie nous permet d'appliquer le théorème limite central, pour obtenir :

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \sqrt{n}(\hat{\theta}_{kn} - \theta_k) \xrightarrow[n\to+\infty]{\mathscr{L}/\mathbb{P}_\theta} \mathcal{N}(0, \mathrm{Var}_\theta(\mathbb{1}_{X_1=k})) = \mathcal{N}(0, \theta_k(1-\theta_k))$$

□

3 Densité de probabilité estimée et comparaison avec la loi uniforme

$\hat{p_n}(\{k\})$: Densité de probabilité estimée
$p_U(\{k\})$: Densité de probabilité de la loi uniforme

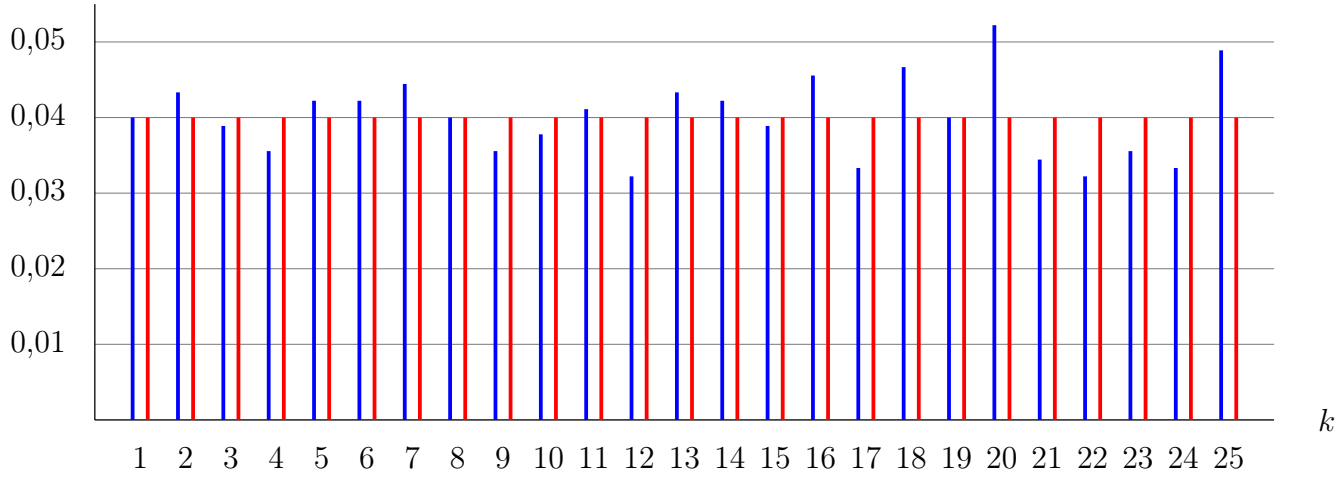

FIGURE 1 – Diagrammes en bâtons de la distribution de probabilité estimée des différentes natures numérotées selon σ ainsi que de la densité de probabilité de la loi uniforme sur $[\![1;25]\!]$.

On observe tout de même que :

$$\sup_{1\leqslant k\leqslant 25} |\hat{p_n}(\{k\}) - p_U(\{k\})| = |\hat{p_n}(\{20\}) - p_U(\{20\})| \approx 0,01222.$$

Ainsi, si notre modèle est juste et bien posé, alors notre taille d'échantillon de 900 semble trop faible pour pouvoir observer nettement une convergence de notre loi vers la loi uniforme sur $[\![1;25]\!]$.

4 Fonction de répartition empirique et comparaison avec la loi uniforme

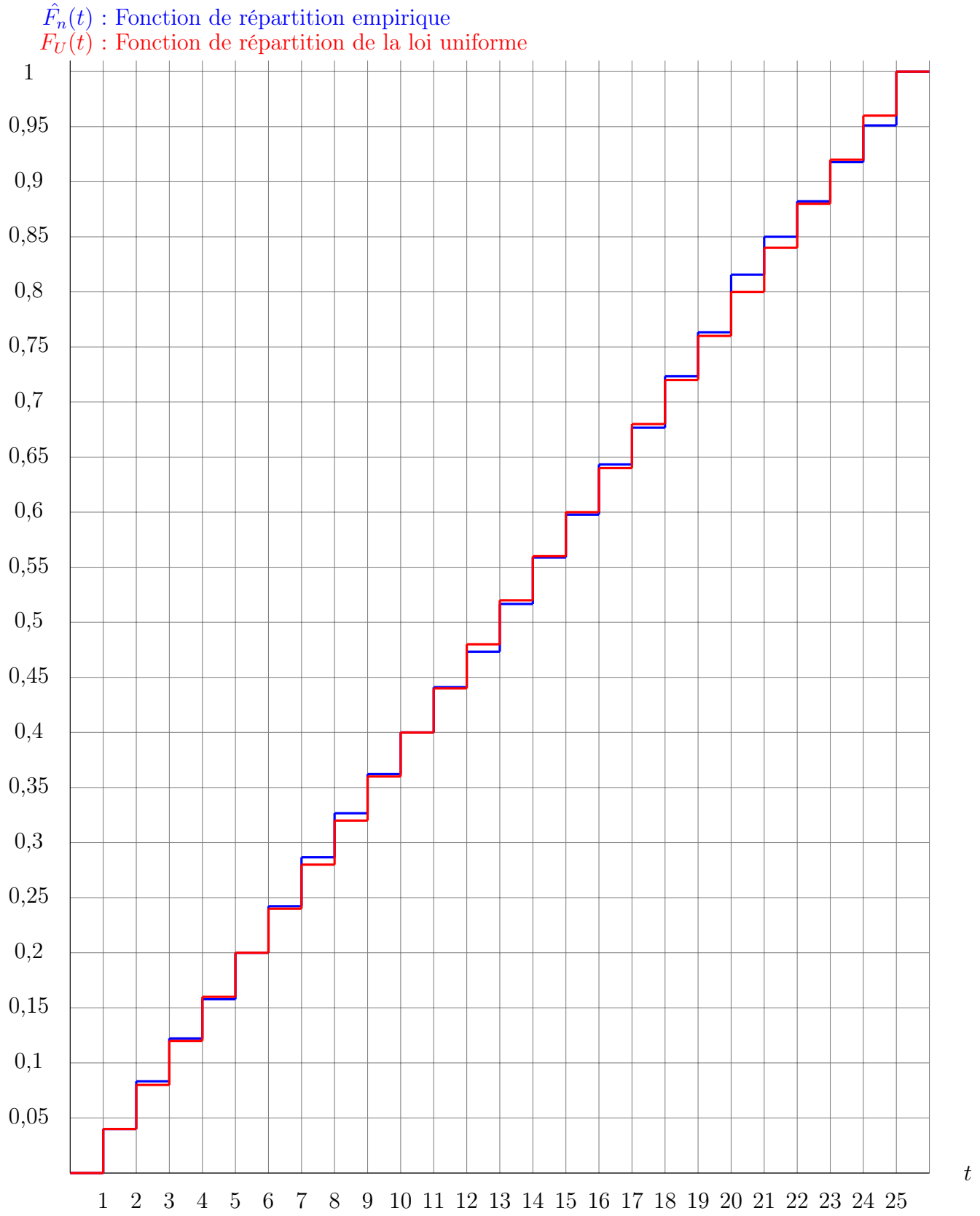

FIGURE 2 – Courbes représentatives de la fonction de répartition empirique des natures numérotées selon σ ainsi que de celle de la loi uniforme sur $[\![1; 25]\!]$.

Pour les fonctions de répartition, on a :

$$\sup_{t\geqslant 0}|\hat{F}_n(t)-F_U(t)| = |\hat{F}_n(20)-F_U(20)| \approx 0,01556.$$

Là encore, la convergence de notre loi vers la loi uniforme ne se voit pas tout à fait nettement pour notre taille d'échantillon.

5 Estimation de nos paramètres par intervalles de confiance

5.1 Construction d'intervalles de confiance asymptotiques

Dans cette section, nous utiliserons le caractère asymptotiquement normal des estimateurs $\hat{\theta}_{kn}$ pour $k \in [\![1;25]\!]$ pour obtenir construire des intervalles de confiance asymptotiques de nos paramètres.

5.1.1 Première méthode : Lemme de Slutsky

Une première méthode classique pour estimer θ_k par intervalles de confiance asymptotique pour $k \in [\![1;25]\!]$ est d'utiliser le lemme de Slutsky pour se ramener à une loi limite ne dépendant pas de θ_k. En effet, la forte consistance de notre estimateur $\hat{\theta}_{kn}$, couplé à la continuité de l'application

$$\begin{array}{ccc}]0,1[& \longrightarrow & \mathbb{R}^{+*} \\ u & \longmapsto & \sqrt{u(1-u)} \end{array}$$

nous donne :

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \sqrt{\hat{\theta}_{kn}(1-\hat{\theta}_{kn})} \xrightarrow[n\to+\infty]{\mathbb{P}_\theta\text{-p.s.}} \sqrt{\theta_k(1-\theta_k)}.$$

En particulier, il y a également convergence en probabilité sous $\mathbb{P}_\theta$ vers la constante $\sqrt{\theta_k(1-\theta_k)} > 0$. Ainsi, le lemme de Slutsky nous donne :

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \left(\sqrt{n}(\hat{\theta}_{kn}-\theta_k), \sqrt{\hat{\theta}_{kn}(1-\hat{\theta}_{kn})}\right) \xrightarrow[n\to+\infty]{\mathscr{L}/\mathbb{P}_\theta} \left(Z, \sqrt{\theta_k(1-\theta_k)}\right)$$

où Z est une variable aléatoire suivant la loi $\mathscr{N}(0,\theta_k(1-\theta_k))$. De plus, par continuité de l'application

$$\begin{array}{ccc} \mathbb{R}\times\mathbb{R}^* & \longrightarrow & \mathbb{R} \\ (x,y) & \longmapsto & \frac{x}{y} \end{array}$$

et par le *continuous mapping theorem*, on a :

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \sqrt{\frac{n}{\hat{\theta}_{kn}(1-\hat{\theta}_{kn})}}(\hat{\theta}_{kn}-\theta_k) \xrightarrow[n\to+\infty]{\mathscr{L}/\mathbb{P}_\theta} \mathscr{N}(0,1).$$

On a alors, en notant Φ la fonction de répartition de la loi normale centrée réduite, et en prenant $\alpha \in]0,1[$:

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \mathbb{P}_\theta\left(\sqrt{\frac{n}{\hat{\theta}_{kn}(1-\hat{\theta}_{kn})}}(\hat{\theta}_{kn}-\theta_k) \in \left[-\Phi^{-1}\left(1-\frac{\alpha}{2}\right); \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]\right) \xrightarrow[n\to+\infty]{} 1-\alpha.$$

i.e.

$$\boxed{\mathbb{P}_\theta\left(\theta_k \in \left[\hat{\theta}_{kn} - \sqrt{\frac{\hat{\theta}_{kn}(1-\hat{\theta}_{kn})}{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right); \hat{\theta}_{kn} + \sqrt{\frac{\hat{\theta}_{kn}(1-\hat{\theta}_{kn})}{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]\right) \xrightarrow[n\to+\infty]{} 1-\alpha.}$$

Ainsi, l'intervalle $\left[\hat{\theta}_{kn} - \sqrt{\frac{\hat{\theta}_{kn}(1-\hat{\theta}_{kn})}{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right); \hat{\theta}_{kn} + \sqrt{\frac{\hat{\theta}_{kn}(1-\hat{\theta}_{kn})}{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]$ constitue un intervalle de confiance asymptotique de θ_k de niveau $1-\alpha$ pour $k \in [\![1;25]\!]$ et $\alpha \in]0;1[$ fixés. Prenons $\alpha = 0,05$ pour obtenir un niveau de confiance classique de 95%. En notant $A_{k,n}$ et $B_{k,n}$ les bornes inférieure et supérieure respectivement de cet intervalle, on a :

k	$A_{k,n}(\omega)$	$B_{k,n}(\omega)$
1 (Hardi)	$0,02719$	$0,05281$
2 (Solo)	$0,03003$	$0,05664$
3 (Rigide)	$0,02625$	$0,05152$
4 (Mauvais)	$0,02345$	$0,04766$
5 (Brave)	$0,02908$	$0,05536$
6 (Assuré)	$0,02908$	$0,05536$
7 (Docile)	$0,03098$	$0,05791$
8 (Malin)	$0,02719$	$0,05281$
9 (Lâche)	$0,02345$	$0,04766$
10 (Relax)	$0,02532$	$0,05023$
11 (Modeste)	$0,02813$	$0,05409$
12 (Doux)	$0,02068$	$0,04376$
13 (Pudique)	$0,03003$	$0,05664$
14 (Foufou)	$0,02908$	$0,05536$
15 (Discret)	$0,02625$	$0,05152$
16 (Calme)	$0,03193$	$0,05918$
17 (Gentil)	$0,02160$	$0,04507$
18 (Prudent)	$0,03288$	$0,06045$
19 (Bizarre)	$0,02719$	$0,05281$
20 (Malpoli)	$0,03768$	$0,06676$
21 (Timide)	$0,02252$	$0,04636$
22 (Pressé)	$0,02068$	$0,04376$
23 (Jovial)	$0,02345$	$0,04766$
24 (Naïf)	$0,02160$	$0,04507$
25 (Sérieux)	$0,03480$	$0,06298$

Les résultats de $A_{k,n}(\omega)$ ont été arrondis par défaut à 10^{-5} près tandis que les résultats de $B_{k,n}(\omega)$ ont été arrondis par excès à 10^{-5} près.

5.1.2 Deuxième méthode : Méthode Delta et fonction arcsinus

Une deuxième méthode pour construire un intervalle de confiance asymptotique pour θ_k pour $k \in [\![1;25]\!]$ est de se ramener à la loi normale centrée réduite à la limite grâce à la méthode Delta. Pour cela, on cherche alors une fonction

$$g \quad : \quad]0;1[\quad \longrightarrow \quad \mathbb{R}$$

de classe $\mathscr{C}^1$ sur $\mathbb{R}$ vérifiant :

$$\forall x \in]0;1[, \quad g'(x) = \frac{1}{\sqrt{x(1-x)}}$$

afin de standardiser la loi normale à la limite.

Rappelons quelques propriétés sur la fonction arcsin. On a que $\arcsin \in \mathscr{C}^\infty(]-1;1[)$ et :

$$\forall x \in]-1,1[, \quad \arcsin'(x) = \frac{1}{\sqrt{1-x^2}} = \frac{1}{\sqrt{(1-x)(1+x)}}.$$

Il semblerait alors qu'une composition par une simple fonction affine puisse donner le résultat escompté. Nous pouvons aisément passer de l'intervalle $]-1;1[$ à l'intervalle $]0;1[$ par la fonction affine suivante :

$$\begin{array}{ccc}]0;1[& \longrightarrow &]-1;1[\\ x & \longmapsto & 2x-1. \end{array}$$

Prenons alors :

$$\begin{array}{rccc} g & : &]0;1[& \longrightarrow & \mathbb{R} \\ & & x & \longmapsto & \arcsin(2x-1). \end{array}$$

On a alors, par composition, $g \in \mathscr{C}^\infty(]0;1[)$ et :

$$\forall x \in]0;1[, \quad g'(x) = \frac{2}{\sqrt{(1-2x+1)(1+2x-1)}} = \frac{2}{\sqrt{2x(2-2x)}} = \frac{1}{\sqrt{x(1-x)}}.$$

On a alors, en appliquant la méthode Delta :

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \sqrt{n}\left(g(\hat{\theta}_{kn}) - g(\theta_k)\right) \xrightarrow[n\to+\infty]{\mathscr{L}/\mathbb{P}_\theta} \mathcal{N}(0,1).$$

Ce qui donne, pour $\alpha \in]0;1[$ fixé et pour tout $\theta \in \Theta$ et pour tout $k \in [\![1;25]\!]$:

$$\mathbb{P}_\theta\left(\sqrt{n}\left(\arcsin\left(2\hat{\theta}_{kn}-1\right) - \arcsin\left(2\theta_k - 1\right)\right) \in \left[-\Phi^{-1}\left(1-\frac{\alpha}{2}\right); \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]\right) \xrightarrow[n\to+\infty]{} 1-\alpha.$$

d'où, par stricte croissance de la fonction sin sur $\left]0;\frac{\pi}{2}\right[$:

$$\boxed{\mathbb{P}_\theta(\theta_k \in C_{k,n}(X)) \xrightarrow[n\to+\infty]{} 1-\alpha,}$$

où :

$$C_{k,n}(X) = \left[\tilde{A}_{k,n}; \tilde{B}_{k,n}\right]$$

avec :

$$\boxed{\tilde{A}_{k,n} = \frac{1}{2}\left(1 + \sin\left(\arcsin\left(2\hat{\theta}_{kn}-1\right) - \frac{1}{\sqrt{n}}\Phi\left(1-\frac{\alpha}{2}\right)\right)\right).}$$

et :

$$\boxed{\tilde{B}_{k,n} = \frac{1}{2}\left(1 + \sin\left(\arcsin\left(2\hat{\theta}_{kn}-1\right) + \frac{1}{\sqrt{n}}\Phi\left(1-\frac{\alpha}{2}\right)\right)\right).}$$

De même, en prenant $\alpha = 0,05$ pour avoir un niveau de confiance de 95%, on a :

k	$\tilde{A}_{k,n}(\omega)$	$\tilde{B}_{k,n}(\omega)$
1 (Hardi)	$0,02818$	$0,05378$
2 (Solo)	$0,03101$	$0,05761$
3 (Rigide)	$0,02725$	$0,05250$
4 (Mauvais)	$0,02445$	$0,04864$
5 (Brave)	$0,03006$	$0,05633$
6 (Assuré)	$0,03006$	$0,05633$
7 (Docile)	$0,03196$	$0,05888$
8 (Malin)	$0,02818$	$0,05378$
9 (Lâche)	$0,02445$	$0,04864$
10 (Relax)	$0,02631$	$0,05122$
11 (Modeste)	$0,02912$	$0,05506$
12 (Doux)	$0,02169$	$0,04475$
13 (Pudique)	$0,03101$	$0,05761$
14 (Foufou)	$0,03006$	$0,05633$
15 (Discret)	$0,02725$	$0,05249$
16 (Calme)	$0,03291$	$0,06014$
17 (Gentil)	$0,02260$	$0,04605$
18 (Prudent)	$0,03386$	$0,06141$
19 (Bizarre)	$0,02818$	$0,05378$
20 (Malpoli)	$0,03865$	$0,06771$
21 (Timide)	$0,02353$	$0,04735$
22 (Pressé)	$0,02169$	$0,04475$
23 (Jovial)	$0,02445$	$0,04864$
24 (Naïf)	$0,02260$	$0,04605$
25 (Sérieux)	$0,03577$	$0,06393$

Là encore, les résultats de $\tilde{A}_{k,n}(\omega)$ ont été arrondis par défaut à 10^{-5} près tandis que les résultats de $\tilde{B}_{k,n}(\omega)$ ont été arrondis par excès à 10^{-5} près.

On remarque que cette fois-ci les intervalles ne sont pas centrés autour de $\hat{\theta}_{kn}$, mais sont, pour certains, plus resserrés, bien que la différence soit minime (de l'ordre de 10^{-5}, tandis que les proportions estimées sont de l'ordre de 10^{-2}). Par exemple :

$$\tilde{B}_{1,n}(\omega) - \tilde{A}_{1,n}(\omega) \approx 0,02560,$$

tandis que

$$B_{1,n}(\omega) - A_{1,n}(\omega) \approx 0,02562.$$

5.2 Construction d'intervalles de confiance non-asymptotiques

Dans cette section, nous utiliserons le fait que les variables aléatoires $(\mathbb{1}_{X_i=k})_{1\leqslant i\leqslant n}$ suivent une loi de Bernoulli $\mathscr{B}(\theta_k)$ pour en déduire des intervalles de confiance non-asymptotiques grâce à l'inégalité de Hœffding. En effet :

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!],\ \forall i \in [\![1;n]\!], \quad 0 \leqslant \mathbb{1}_{X_i=k} \leqslant 1 \quad \mathbb{P}_\theta\text{-p.s..}$$

Ainsi, l'inégalité de Hœffding s'applique, et on obtient, pour tout $x > 0$:

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \mathbb{P}_\theta\left(n\hat{\theta}_{kn} - n\theta_k \geqslant x\right) = \mathbb{P}_\theta\left(\sum_{i=1}^{n} \mathbb{1}_{X_i=k} - \mathbb{E}_\theta\left(\sum_{i=1}^{n} \mathbb{1}_{X_i=k}\right) \geqslant x\right) \leqslant \exp\left(-\frac{2x^2}{n}\right).$$

L'inégalité de Hœffding s'applique également aux variables aléatoires $(1 - \mathbb{1}_{X_i=k})_{1\leqslant i\leqslant n}$ qui vérifient également :

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!],\ \forall i \in [\![1;n]\!], \quad 0 \leqslant 1 - \mathbb{1}_{X_i=k} \leqslant 1 \quad \mathbb{P}_\theta\text{-p.s..}$$

Ainsi, on obtient, pour tout $x > 0$:

$$\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \mathbb{P}_\theta\left(-n\hat{\theta}_{kn} + n\theta_k \geqslant x\right) \leqslant \exp\left(-\frac{2x^2}{n}\right).$$

Ce qui donne :

$$\forall x > 0,\ \forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \mathbb{P}_\theta\left(\left|n\hat{\theta}_{kn} - n\theta_k\right| \geqslant x\right) \leqslant 2\exp\left(-\frac{2x^2}{n}\right),$$

et donc :

$$\forall x > 0,\ \forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \mathbb{P}_\theta\left(\left|\hat{\theta}_{kn} - \theta_k\right| \geqslant x\right) \leqslant 2\exp\left(-2nx^2\right).$$

Pour obtenir alors un intervalle de confiance non-asymptotique de niveau $1-\alpha$ pour $\alpha \in]0;1[$, on choisit $x > 0$ tel que :

$$2\exp(-2nx^2) = \alpha.$$

i.e.

$$x = \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)}.$$

On obtient alors :

$$\boxed{\forall \theta \in \Theta,\ \forall k \in [\![1;25]\!], \quad \mathbb{P}_\theta\left(\theta_k \in \left[\hat{\theta}_{kn} - \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)}; \hat{\theta}_{kn} + \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)}\right]\right) \geqslant 1 - \alpha.}$$

Ainsi, l'intervalle $\left[\hat{\theta}_{kn} - \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)}; \hat{\theta}_{kn} + \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)}\right]$ constitue un intervalle de confiance non-asymptotique pour l'estimation de θ_k au niveau de confiance $1 - \alpha$. Étant donné que le paramètre θ_k est strictement positif, la borne $\hat{\theta}_{kn} - \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)}$ peut être remplacée par $\left(\hat{\theta}_{kn} - \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)}\right)_+$, où x_+ désigne la partie positive de x.

Dans le cadre de notre expérience, et en prenant $\alpha = 0,05$, on a :

$$\sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)} \approx 0,04527. \quad \text{(résultat arrondi à } 10^{-5} \text{ près)}$$

Notre intervalle de confiance non-asymptotique est donc bien plus large que nos intervalles de confiances asymptotiques ! En effet, en notant $\hat{A}_{k,n}$ et $\hat{B}_{k,n}$ les bornes inférieure et supérieure respectivement de

notre intervalle de confiance non-asymptotique, on obtient :

k	$\hat{A}_{k,n}(\omega)$	$\hat{B}_{k,n}(\omega)$
1 (Hardi)	0	$0,08528$
2 (Solo)	0	$0,08861$
3 (Rigide)	0	$0,08416$
4 (Mauvais)	0	$0,08083$
5 (Brave)	0	$0,08750$
6 (Assuré)	0	$0,05633$
7 (Docile)	0	$0,08972$
8 (Malin)	0	$0,08528$
9 (Lâche)	0	$0,08083$
10 (Relax)	0	$0,08305$
11 (Modeste)	0	$0,08639$
12 (Doux)	0	$0,07750$
13 (Pudique)	0	$0,08861$
14 (Foufou)	0	$0,08750$
15 (Discret)	0	$0,08416$
16 (Calme)	$0,00029$	$0,09083$
17 (Gentil)	0	$0,07861$
18 (Prudent)	$0,00140$	$0,09194$
19 (Bizarre)	0	$0,08528$
20 (Malpoli)	$0,00696$	$0,09750$
21 (Timide)	0	$0,07972$
22 (Pressé)	0	$0,07750$
23 (Jovial)	0	$0,08083$
24 (Naïf)	0	$0,07861$
25 (Sérieux)	$0,00362$	$0,09416$

Comparons alors ces différents intervalles de confiance sur un graphique, pour l'exemple de la nature Malpoli :

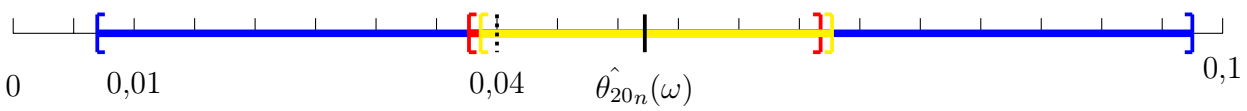

FIGURE 3 – Comparaison graphique des différents intervalles de confiance construits pour la nature Malpoli : Inégalité de Hœffding, Théorème Limite Central + Lemme de Slutsky et Théorème Limite Central + méthode Delta (en jaune).

6 Conclusion : peut-on rejeter l'hypothèse selon laquelle les natures dans pokémon sont distribuées selon une loi uniforme ?

Comme suggéré dans le titre de cette section, tout comme dans l'introduction, nous testons notre hypothèse nulle H_0 : « Les variables aléatoires $(X_i)_{1\leqslant i\leqslant n}$ suivent la loi uniforme sur $[\![1;25]\!]$. » contre l'hypothèse alternative H_1 : « Les variables aléatoires $(X_i)_{1\leqslant i\leqslant n}$ ne suivent pas la loi uniforme sur $[\![1;25]\!]$. ».

6.1 Étape 1 : En considérant les intervalles de confiance construits

Une manière de rejeter l'hypothèse nulle avec grande probabilité est de vérifier si la probabilité $0,04$, correspondant à la probabilité d'un singleton de $[\![1;25]\!]$ sous la loi uniforme, est dans nos intervalles de confiance. Si tel n'est pas le cas, on pourra considérer l'hypothèse H_0 rejetée.

Conclusion : $0,04$ appartient à tous nos intervalles de confiance construits, non-asymptotiques comme asymptotiques. Nous ne pouvons pas rejeter notre hypothèse H_0 avec cette observation.

6.2 Étape 2 : En réalisant le test d'adéquation du χ^2

Rappelons brièvement le principe de ce test dans le cadre de notre exemple : afin de contrôler l'erreur de première espèce (i.e. la probabilité que H_0 soit fausse alors qu'elle est effectivement vraie), on introduit la statistique du χ^2 :

$$T_n = n \sum_{k=1}^{25} \left(\frac{\left(\hat{\theta}_{kn} - 0,04\right)^2}{0,04} \right).$$

On a alors l'alternative suivante :

— Si H_0 est vraie, alors $T_n \xrightarrow[n\to+\infty]{\mathscr{L}} \chi^2(24)$.

— Si H_1 est vraie, alors $T_n \xrightarrow[n\to+\infty]{\text{p.s.}} +\infty$.

Ainsi, en prenant une marge d'erreur $\alpha \in]0;1[$ et le quantile $q_{1-\alpha}$ d'ordre $1-\alpha$ de la loi $\chi^2(24)$, on a, si H_0 est vraie :

$$\mathbb{P}(T_n \geqslant q_{1-\alpha}) \xrightarrow[n\to+\infty]{} \alpha.$$

On considèrera donc l'hypothèse H_0 rejetée si l'on obtient $T_n \geqslant q_{1-\alpha}$, au risque de se tromper de l'ordre de α, si n est assez grand.

Le module `scipy.stats` nous permet d'effectuer directement le test d'adéquation du χ^2 sur Python. En entrant les commandes

```
import scipy.stats as st
ListEffectif = [36, 39, 35, 32, 38, 38, 40, 36, 32, 34, 37, 29, 39, 38, 35, 41, 30, 42, 36, 47,
↪  31, 29, 32, 30, 44]
st.chisquare(ListEffectif) #Effectue le test d'adéquation du chi-2 en prenant comme loi
↪  théorique la loi uniforme.
```

dans Python, on obtient le résultat suivant :

```
Power_divergenceResult(statistic=15.055555555555557, pvalue=0.9191225747213196)
```

Cela veut dire que Python a calculé la statistique du χ^2 en notre ω et il trouve :

$$T_n(\omega) \approx 15,05556.$$

Python, via cette commande, a également calculé la *valeur-p* du χ^2 pour la statistique trouvée, c'est-à-dire la valeur $p \in]0;1[$ telle que $T_n(\omega) = q_{1-p}$, et il trouve :

$$p \approx 0,91912.$$

Conclusion : En se fixant au préalable notre niveau de risque $\alpha = 0,05$, on conclut, étant donné que $q_{1-\alpha} \approx 36,42$ et que $T_n(\omega) < 36,42$, que le test d'adéquation du χ^2 ne nous permet pas non plus de rejeter l'hypothèse selon laquelle les natures dans pokémon sont distribuées selon une loi uniforme. On arrive à la même conclusion avec la *valeur-p* : $0,91912 > \alpha$, donc on peut accepter notre hypothèse H_0.

Annexes

FIGURE 4 – Exemple de statistiques d'un pokémon.

Le pokémon dont les statistiques sont montrées ci-dessus possède la nature Mauvais : il a donc un malus dans sa statistique de défense spéciale et un bonus dans sa statistique d'attaque. Sans ce malus en défense spéciale, le pokémon aurait eu une valeur de défense spéciale égale à 159. En effet :

$$159 \times 0,1 = 15,9$$

et

$$159 - 15,9 = 143,1.$$

Étant donné que les valeurs de statistiques doivent être entières, le jeu tronque ces valeurs. En tronquant $143,1$, on obtient 143. De même, sans le bonus en attaque, le pokémon aurait eu une valeur d'attaque égale à 329. En effet :

$$329 \times 0,1 = 32,9$$

et

$$329 + 32,9 = 361,9.$$

Là encore, en tronquant, on obtient une valeur de 361.