Estimations

Échantillonnage, fluctuations et prises de décisions.

Intervalles de fluctuation

Principe

On considère:

• une **population** $\mathcal{P}$ possédant un **caractère** C de proportion $p_{\text{théorique}}$ dans cette population.

• des **échantillons** de **taille** N de cette population. On note $p_{\text{observée}}$ la proportion du caractère C dans cet échantillon.

On fait varier la taille N de l'échantillon:

• les proportions observées sont distinctes de la proportions théoriques en général

• plus la taille de l'échantillon est grande, et plus la proportions observées se rapprochent de la proportion théorique.

On observe en choisissant un nombre grand d'échantillon de même taille N que:

• il est peu probable que $p_{\text{théorique}} = p_{\text{observée}}$

• 2 échantillons présentent des proportions observées en général distinctes

• les proportions observées pour des échantillons de même taille ont tendance à se trouver autour de la proportion théorique.

L'ensemble des proportions observées pour un ensemble d'échantillon de même taille est appelée la fluctuation d'échantillonnage.

Ces résultats sont traduits par la **loi des grands nombres**.

Pour résumer, l'idée est que ce n'est pas parce que l'on a 1 chance sur 6 de tirer 6 quand on lance un dé à 6 faces que sur 6 lancers on obtiendra forcément un 6.
Par contre en répétant le nombre de lancers, la probabilité de faire un 6 quand on n'en a pas encore fait, augmente.
De plus si on lance un nombre N de fois un même dé, et que l'on répète un autre grand nombre de fois des séries de N lancers, on observera que la proportion de 6 obtenue pour chaque série de N lancers est d'autant plus proche de $\frac{1}{6}$ que le nombre N est grand et d'autre part, les proportions de 6 observées pour chaque série de N sont majoritairement proches de la proportion théorique.

Un intervalle qui contient une majeure partie de ces valeurs est appelé un **intervalle de fluctuation**.

On veut pouvoir estimer a priori des intervalles de fluctuations à une précision donnée.

Dans ce paragraphe, la proportion $p_{\text{théorique}}$ est **connue**.

On note F_N la variable aléatoire qui à un échantillon de taille N associe sa fréquence F_n de la proportion du caractère C dans l'échantillon choisi.

Définition:

Un intervalle de fluctuation de la variable aléatoire F_N au seuil $0,95$ (au seuil de 95%) est un intervalle contenant F_N avec une probabilité supérieure à $0,95$.

■ ***Remarques:***

• Cet intervalle est en pratique choisi de plus petite amplitude possible.

• Il n'y a pas unicité de l'intervalle de fluctuation.

• On peut choisir un intervalle de fluctuation au seuil $1 - \alpha$ % voulu en adaptant la définition.

Intervalles de fluctuations (rappels)

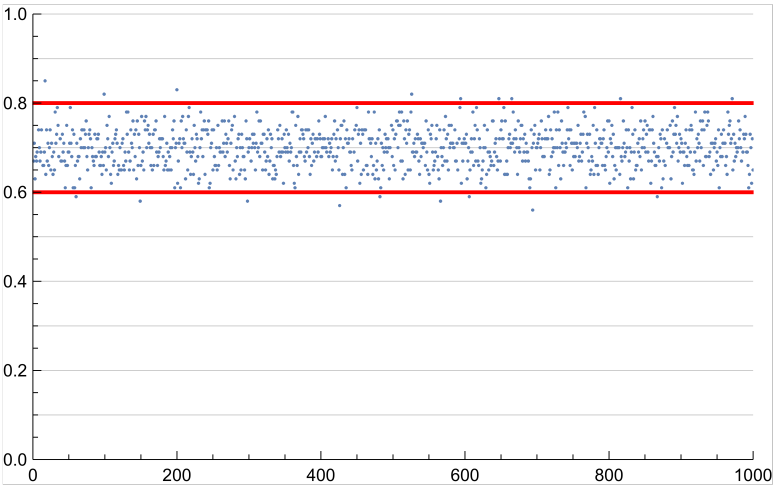
Intervalle de fluctuation (seconde)

Par l'observation et la simulation, on obtient que:

Si $\begin{cases} N \geq 30 \\ 0,2 \leq p_{\text{théorique}} \leq 0,8 \end{cases}$, l'intervalle $\left[p_{\text{théorique}} - \frac{1}{\sqrt{N}}; p_{\text{théorique}} + \frac{1}{\sqrt{N}}\right]$ est un intervalle de fluctuation au seuil 0,95 de la variable aléatoire F_N.

■ ***Illustration:***

On a simulé 1000 échantillons de taille 100, pour une probabilité du caractère C $p_{\text{vraie}} = 0,4$.
Chaque point représente la fréquence observé pour un échantillon, les droites horizontales représentent les bornes de l'intervalle de fluctuation.



Intervalle de fluctuation:[0.6;0.8]
Nombre de tirages en dehors de l'intervalle de fluctuation: 18
Pourcentage de tirages en dehors de l'intervalle de fluctuation: 1.8%

Intervalle de fluctuation (première)

On choisit un individu au hasard dans cette population, supposée suffisamment grande pour que .
Il a donc la probabilité $p_{\text{théorique}}$ d'avoir le caractère C.
On choisit N fois un individu de cette population supposée suffisamment grande pour que le tirage d'un individu puisse être assimilé à un tirage avec remise.
On répète donc N fois de manière indépendante la même épreuve de Bernoulli de paramètre de succès $p_{\text{théorique}}$.
Alors la variable aléatoire X_N qui dénombre le nombre d'individus parmi les N choisis ayant le caractère C suit une loi binomiale $\mathcal{B}(N; p_{\text{théorique}})$.

On définit alors la variable aléatoire $F_N = \dfrac{X_N}{N}$ décrivant la fréquence du caractère C dans un échantillon de taille N.

On a $X_N \in \{0;\ \ldots;\ N\}$ donc $F_N \in \left\{0;\ \dfrac{1}{N};\ \dfrac{2}{N};\ \ldots;\ \dfrac{N}{N} = 1\right\}$.

Attention: F_N ne suit pas une loi binomiale puisque les valeurs ne sont pas entières, mais pour tout entier $k \in \{0;\ \ldots;\ N\}$, on a $p\left(F_N = \dfrac{k}{N}\right) = p(X_N = k) = \binom{N}{k} p_{\text{théorique}}{}^{k}\left(1 - p_{\text{théorique}}\right)^{N-k}$.

On considère un réel $\alpha \in [0; 1]$ (pour fixer les idées $\alpha = 5$).
Alors comme $1 - \alpha \in [0; 1]$, il existe des entiers a et b avec $0 \leq a \leq b \leq n$ tels que $p(a \leq X_N \leq b) \geq 1 - \alpha$ ou encore tels que $p((X \leq a) \cup (X \geq b)) \leq \alpha$.

On en déduit que $p\left(\dfrac{a}{N} \leq F_N \leq \dfrac{b}{N}\right) \geq 1 - \alpha$ et $p\left(\left(F_N \leq \dfrac{a}{N}\right) \cup \left(F_N \geq \dfrac{b}{N}\right)\right) \leq \alpha$.

Ainsi on peut déterminer un intervalle de fluctuation au seuil de α % en utilisant la loi binomiale (en particulier la fonction inverse-binomiale de la calculatrice).

- ***Méthode:***

Au seuil de 5%:
On détermine à l'aide de la fonction "InverseBinomial" les entiers a et b tels que $p(X_N \leq a) < 0,025$ (2, 5 %) et $p(X_N \leq b) > 0,975$ (donc $p(X_N > b) < 0,025$.
Alors $p\left(\dfrac{a}{N} \leq F_N \leq \dfrac{b}{N}\right) = p(a \leq X_N \leq b) \geq 0,95$.
On a trouvé un intervalle de fluctuation de la variable aléatoire $F_N$au seuil 0, 95.

Intervalle de fluctuation asymptotique

Théorème de Moivre-Laplace (version simplifiée)

Soit $p \in [0; 1]$.
Soit (X_n) une suite de variable aléatoire suivant les lois binomiales $\mathcal{B}(n, p)$.

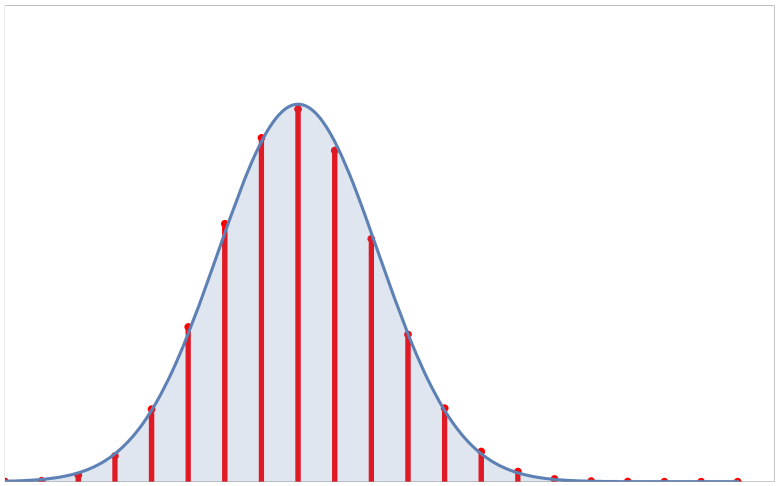
Alors la suite $(Z_n) = \left(\dfrac{X_n - n\,p}{\sqrt{n\,p(1-p)}}\right)$ des variables aléatoires centrées réduites associées à la suite (X_n) converge vers une variable aléatoire qui suit une loi normale $\mathcal{N}(0; 1)$.

Autrement dit, pour un entier n assez grand, si X_n suit la loi $\mathcal{B}(n; p)$ alors pour tous réels a et b,

$$p(a \leq X_n \leq b) \approx p\left(\frac{a - n\,p}{\sqrt{n\,p\,(1-p)}} \leq Z_n \leq \frac{b - n\,p}{\sqrt{n\,p(1-p)}}\right) \approx p\left(\frac{a - n\,p}{\sqrt{n\,p\,(1-p)}} \leq \mathcal{N}(0; 1) \leq \frac{b - n\,p}{\sqrt{n\,p(1-p)}}\right).$$

On peut donc utiliser la loi normale pour obtenir des approximations de calcul de probabilités quand on travaille sur une population assez grande, plutôt que d'utiliser les résultats sur la loi binomiale.

■ ***Illustration:***



Définition:

Un intervalle de fluctuation asymptotique de la variable aléatoire F au seuil 0,95 est un intervalle déterminé à partir de p et de N et qui contient F avec une probabilité d'autant plus proche de 0, 95 que N est grand.

■ ***Remarques:***

On dit asymptotique pour indiquer la notion de limites: "N devient grand ".

Le théorème des valeurs intermédiaires appliqué à la fonction $x \mapsto \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt = p(X \leq x)$ lorsque la variable aléatoire X suit une loi $\mathcal{N}(0; 1)$ permettent de démontrer le résultat qui suit. On pourrait choisir le seuil qui nous convient.

Intervalle de fluctuation asymptotique (terminale)

Un intervalle de fluctuation asymptotique au seuil 0, 95 de la variable aléatoire fréquence F est l'intervalle:

$$\left[p_{\text{théorique}} - 1,96 \frac{\sqrt{p_{\text{théorique}}(1 - p_{\text{théorique}})}}{\sqrt{N}}; p_{\text{théorique}} + 1,96 \frac{\sqrt{p_{\text{théorique}}(1 - p_{\text{théorique}})}}{\sqrt{N}}\right].$$

Prise de décision

Situation:

Dans ce paragraphe, on suppose que la **proportion** $p_{\text{théorique}}$ du caractère étudié **est supposée être égale à** p.

La **prise de décision** consiste, à partir d'un échantillon de taille N, à valider ou cette hypothèse-faite sur la proportion p.

Prise de décision: l'hypothèse de conformité.

On suppose que $n \geq 30$, $n\,p \geq 5$ et $n(1 - p) \geq 5$.

On fait l'hypothèse, dite hypothèse nulle ou hypothèse de **conformité**, H_0 : $p = p_{\text{théorique}}$. (la proportion p est notre proposition et $p_{\text{théorique}}$ est la vraie proportion).

Règle:

Si la fréquence observée $f_{obs,n} \notin I_f = \left[p - 1,96\,\frac{\sqrt{p(1-p)}}{\sqrt{N}}; p + 1,96\,\frac{\sqrt{p(1-p)}}{\sqrt{N}}\right]$, on rejette l'hypothèse H_0, sinon, on ne la rejette pas.

Autrement dit la règle de prise de décision consiste à:

• si la fréquence $f_{\text{observée}}$ appartient à l'intervalle de fluctuation asymptotique au seuil 0, 95, on accepte l'hypothèse faite sur la proportion $p_{\text{théorique}}$.

• si la fréquence $f_{\text{observée}}$ n'appartient pas à l'intervalle de fluctuation asymptotique au seuil 0, 95, on rejette l'hypothèse faite sur la proportion p avec un risque de 5 % de se tromper.

■ ***Remarques:***

• D'après le théorème précédent, la probabilité de rejeter **à tort** l'hypothèse $H_0 : p = p_0$ est environ égale à 0,05. Le seuil de décision correspond à ce risque.

• **Attention.** « Ne pas rejeter l'hypothèse » est différent de « l'hypothèse est vraie ». La règle de prise de décision donnée ici est une règle de **non-conformité** avec un risque d'erreur de 5%.
Dans le cas où on accepte l'hypothèse faite sur la proportion $p_{\text{théorique}}$, le risque d'erreur n'est pas quantifié.

Exemple:

Pour situer, Imaginons que l'on veuille tester si un dé est bien équilibré ou truqué sur le numéro 6.

On fait l'hypothèse de conformité $H_0 : p_{\text{théorique}} = \frac{1}{6}$.

On effectue un nombre de lancers de dés suffisants pour respecter les conditions d'application de l'intervalle de fluctuations asymptotiques.
On calcule alors la fréquence observée f_{obs} de 6.
Si $f_{\text{obs}} \in I_f$ alors on acceptera l'hypothèse que le dé est équilibré sur le 6 au seuil 0, 95.
Sinon on rejettera l'hypothèse et on considérera le dé truqué au risque d'erreur de 5%.
En effet, le dé peut être équilibré et pourtant ne sortir aucun 6 sur un grand nombre de lancers, c'est rare, mais ça peut arriver.
Si le tirage correspond à ce cas, on rejette l'hypothèse à tort.

Intervalle de confiance

Situation:

Dans ce paragraphe, on suppose que la **proportion** $p_{\text{théorique}}$ du caractère étudié **est inconnue**.

Intervalle de confiance

Définition

Un **intervalle de confiance** pour une proportion p au niveau de confiance 0, 95 est la réalisation à partir d'un échantillon, d'un **intervalle aléatoire** contenant la proportion p avec une probabilité supérieure ou égale à 0, 95.

■ ***Remarques:***

• On écrit **un** intervalle de confiance. Il n'y a pas unicité de l'intervalle de confiance au niveau de confiance 95 %.

Propriété

Un intervalle de confiance au niveau de confiance 95 % est l'intervalle défini par $\left[f_{\text{observée}} - \frac{1}{\sqrt{N}}; f_{\text{observée}} + \frac{1}{\sqrt{N}}\right]$, où $f_{\text{observée}}$ est la fréquence observée du caractère étudié sur un **échantillon de taille** N avec $N \geq 30$, $N \times f_{\text{observée}} \geq 5$ et $N \times (1 - f_{\text{observée}}) \geq 5$.

■ ***Remarques:***

• Les conditions d'application ne peuvent être appliquées à la proportion $p_{\text{théorique}}$ théorique, inconnue. On montre que les vérifier avec la fréquence $f_{\text{observée}}$ suffit.

Exemple

On veut estimer le biais d'une pièce de monnaie.
On réalise un grand nombre N de lancers et on note la fréquence f_{obs} de «Pile».
Alors on pourra affirmer au niveau de confiance 0, 95 que la pièce donne «Pile» pour une probabilité p appartenant à l'intervalle $\left[f_{\text{obs}} - \frac{1}{\sqrt{N}}; f_{\text{obs}} + \frac{1}{\sqrt{N}}\right]$.