Université de Lille 1 SCIENCES ET TECHNOLOGIES

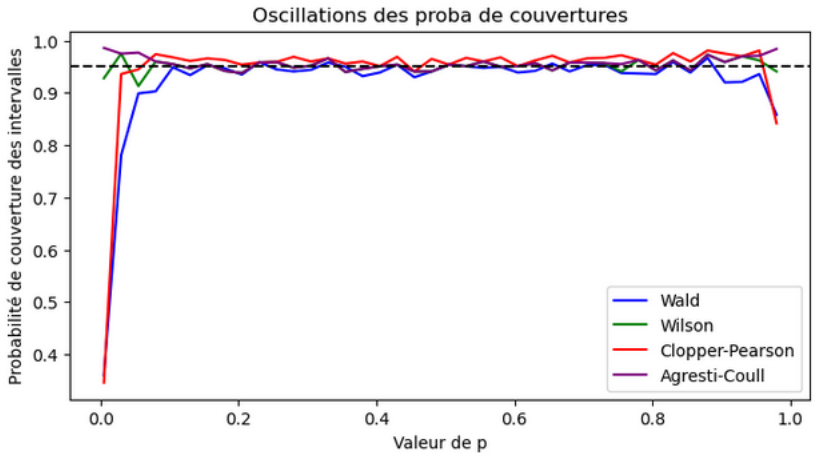
Licence Informatique Appliquées aux Science Humains et Sociales (MIASHS)

Option :

Mathématiques statistiques et informatique décisionnelle (MSID)

PROJET DE FIN DE CYCLE de LICENCE

Comparaison réelle des formules d'intervalles de confiance pour une proportion



source obtenue sous python à n = 40, voir chap.3

Travail réalisé par :
Ahmed ali HADAD
2022-2023

Sous l'encadrement de :
Mr **Baba THIAM** maître de conférences
Equipe de probabilité statistique Laboratoire Paul Painlevé

Résumé

Les intervalles de confiance pour les proportions sont couramment utilisés dans de nombreux domaines, tels que la santé publique, la recherche marketing, la sociologie et la politique, pour évaluer les taux de réussite, les préférences des consommateurs, les opinions publiques et les résultats des élections. Ils peuvent également être utilisés pour comparer les proportions entre différents groupes ou pour évaluer les changements dans les proportions au fil du temps.
En somme, l'intervalle de confiance d'une proportion est une mesure clé pour comprendre l'incertitude associée à l'estimation de la proportion. Il est essentiel pour la prise de décision et pour évaluer la fiabilité des résultats obtenus à partir d'un échantillon. Beaucoup sont les chercheurs qui ont passé du temps à l'étude de l'intervalle de confiance d'une proportion. On se limite à quatre fameux méthodes, nous les avons étudié avec soin et voici les résultats qu'on a trouvé :

1. **Méthode de Wald** : Cette méthode approxime la loi binomiale à la loi normale via le TCL auto-normalisé. Elle utilise l'estimation de la proportion p pour calculer l'intervalle de confiance. Cependant, cette méthode peut donner des intervalles de confiance qui ne recouvrent pas la vraie proportion avec une probabilité inférieure au niveau nominal.
2. **Méthode de Wilson** : Cette méthode part sur l'idée de WALD en corrigeant le centrage de l'intervalle. Elle utilise également une correction pour éviter les intervalles de confiance qui dépassent les limites de [0,1]. Cette méthode donne généralement de bons résultats, avec des probabilités de recouvrement qui fluctuent tout au tour du niveau nominal.
3. **Méthode d'Agresti-Coull** : Cette méthode est une amélioration de la méthode de Wilson. Elle ajoute un nombre fixe de succès et d'échecs artificiels à l'échantillon pour augmenter la précision de l'estimation de la proportion. Cette méthode peut donner des résultats similaires à la méthode de Wilson, mais elle présenter des oscillations plus fortes dans les probabilités de recouvrement qui beaucoup sont au dessus du niveau nominal.
4. **Méthode de Clopper-Pearson** : Cette méthode utilise la distribution exacte de la loi de probabilité binomiale via la loi bêta pour calculer l'intervalle de confiance. Elle garantit que la vraie proportion est incluse dans l'intervalle de confiance avec une probabilité de couverture égale ou supérieure au niveau nominal. Cependant, cette méthode peut donner des intervalles de confiance plus larges que les autres méthodes.

Il est important de noter que le choix de la méthode d'intervalle de confiance dépend de la situation particulière et des objectifs de l'analyse. Nous verrons ainsi que le choix de la taille de l'échantillon à un impact sur les résultats, plus la taille de l'échantillon est grande plus l'intervalle de confiance est précis. Nous verrons chaque expérience simuler en choisissant deux taille d'échantillon n=40 et n=100 pour bien remarquer l'effet produit par la taille de l'échantillon.

Table des matières

Chapitre 1

Généralité et Rappels

Tout au long de ce travail, nous allons utiliser des termes spécifiques et appliquer des concepts qui ne seront pas explicitement définis. Il est donc judicieux de présenter quelques notions et définitions qui seront utilisées dans les prochains chapitres.

1.1 Définition

Définition 1. ***Loi binomiale :***
La loi binomiale, de paramètres n et p, est la loi de probabilité discrète d'une variable aléatoire X somme des $(X_i)_{1\leqslant i\leqslant n}$ variables aléatoires identiquement distribués par la loi bernouilli de paramètre p dont la ***fonction de masse*** *est donnée par :*

$$P(X{=}k) = \binom{n}{k} p^k q^{n-k},\ \forall \quad k = 0, 1, \cdots, n,$$

d'espérance np et de variance $np(1-p)$. Cette loi permet de déterminer le nombre de fois on a succès dans les n expérience de bernouilli.

L'outil principal que nous utiliserons pour déterminer la proportion de succès p est l'intervalle de confiance. Ces intervalles de confiance sont obtenus à partir de ce que l'on appelle les tests statistiques..

Définition 2. ***Test statistique [6] :***
Soit $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_n})$ un échantillon des variables aléatoires de réalisation $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n})$. Un ***test statistique*** *d'hypothèse H_0, dite hypothèse nulle, contre l'hypothèse H_1, dite hypothèse alternative, est une fonction Φ telle que :*

$$\begin{cases} \Phi(x) = 0 & \text{correspond à conserver} \quad H_0 \\ \Phi(x) = 1 & \text{correspond à conserver} \quad H_1 \end{cases}$$

En générale, Φ est de la forme :

$$\Phi(x) = 1_{h(x)\in\mathcal{R}}$$

où **h(x)** est appelé la **statistique de test** et $\mathcal{R}$ est appelé la **zone du rejet**. Prendre une décision à partir d'un test statistique n'élimine pas la possibilité d'erreurs. Il est donc important de connaître la puissance du test pour évaluer si la décision prise est justifiée ou non. un tel problématique incite à prendre en compte les calculs d'erreur et de puissance du test :

Définition 3. ***Erreur et puissance du test [5] :***
Soit Φ *un test statistique (* $\omega, F, (P_\theta)_{\theta \in H}$ *) :*

$$H = H_0 \cup H_1$$

L'erreur de première espèce mesure le niveau d'erreur que l'on commet en rejetant à tort ***l'hypothèse nulle*** H_0 *et en choisissant à la place* ***l'hypothèse alternative H1****, alors que* H_0 *est en fait vraie. Cette erreur est le niveau de significativité de l'hypothèse, donné par :*

$$\begin{aligned} \alpha : H_0 &\longrightarrow [0,1] \\ \theta &\longmapsto \mathbf{P}_\theta[\Phi(x) = 1] \end{aligned}$$

L'erreur du ***deuxième espèce*** *mesure le niveau qu'on ne se trompe pas en choisissant l'****hypothèse alternative*** H_1 *sachant que l'****hypothèse nulle*** H_0 *est fausse, donné par :*

$$\begin{aligned} \beta : H_1 &\longrightarrow [0,1] \\ \theta &\longmapsto \mathbf{P}_\theta[\Phi(x) = 0] \end{aligned}$$

La ***puissance*** *du test mesure le niveau qu'on se trompe pas en rejetant* H_0 *au profit de* H_1*, donnée par :*

$$\begin{aligned} \gamma : H_1 &\longrightarrow [0,1] \\ \theta &\mathbf{P}_\theta[\Phi(x) = 1] \\ &\longmapsto 1 - \beta(\theta) \end{aligned}$$

Lorsque l'on effectue un test statistique, l'attention se porte surtout sur le niveau de signification, c'est-à-dire le seuil auquel on doit rejeter l'hypothèse nulle. Pour une taille d'échantillon d'observation donnée, on peut ainsi déterminer de manière précise quelle hypothèse doit être conservée ou rejetée :

Définition 4. ***P-value ou probabilité critique [5] :***
Soit un test statistique Φ_α de niveau α construit dans un échantillon d'observation $\mathbf{X}$*, on appelle* ***probabilité critique*** *du test le niveau de confiance donné par :*

$$\begin{aligned} \mathbf{P} : X &\longrightarrow [0,1] \\ x &\longmapsto \mathbf{P}(\mathbf{x}) = inf\{\alpha \in [0,1] \quad tel\ que \quad \Phi_\alpha(x) = 1\} \end{aligned}$$

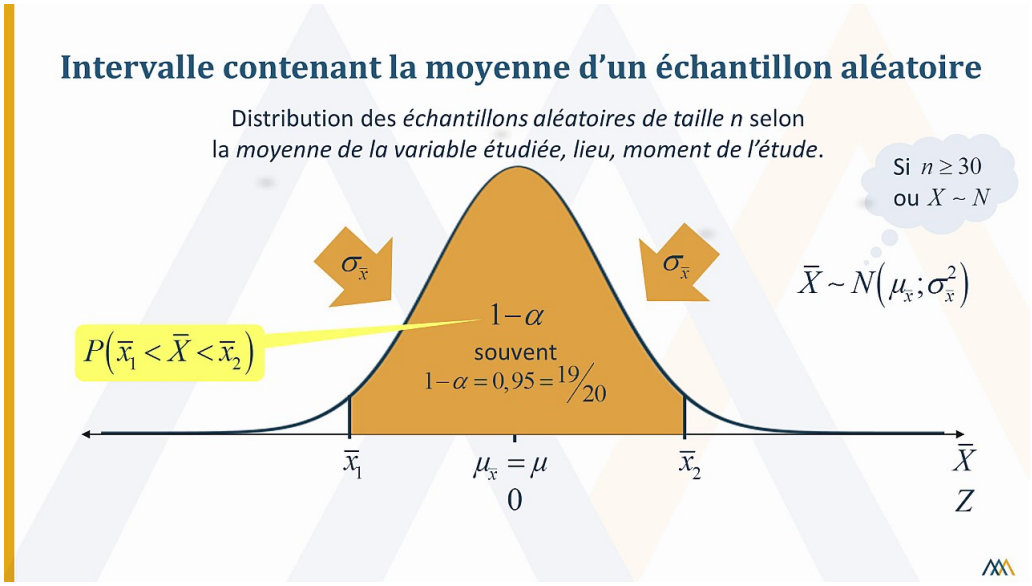
C'est le plus petit niveau pour lequel on rejette l'hypothèse nulle H_0, ainsi :

1. *si $\alpha < \mathbf{P}(\mathbf{x})$ on conserve H_0*

2. *si $\alpha \geq \mathbf{P}(\mathbf{x})$ on conserve H_1*

1.2 Intervalle de confiance

Un intervalle de confiance est construit à partir de deux hypothèses en opposition : **l'hypothèse nulle** H_0 et **l'hypothèse alternative** H_1. L'intervalle de confiance représente une plage de valeurs dans laquelle on peut être confiant que la valeur d'un paramètre μ, qui est encore inconnu, se situe avec un certain niveau de confiance $1-\alpha$.



Source : image prise sur internet

Définition 5. ***Intervalle de confiance [6] :***
Un ***intervalle de confiance*** *de niveau $1-\alpha$ pour une quantité β, construit à partir d'un échantillon $X = (X_1, X_2, \cdots, X_n)$, est un intervalle aléatoire $I(X_1, X_2, \cdots, X_n)$ ne dépend pas de la quantité β et tel que :*

$$\forall \quad \beta, \quad P[\beta \in I(X_1, X_2, \cdots, X_n)] \geqslant 1-\alpha$$

1.3 Convergence en loi

L'un des outils employés tout au long de ce travail est la **convergence en loi**. Ce type de convergence est largement utilisé en statistique pour étudier le comportement asymptotique de séquences de variables aléatoires dont la loi est inconnue. L'objectif principal est de démontrer que ces variables aléatoires convergent en loi vers une autre variable aléatoire dont la loi est connue.

Définition 6. ***Convergence en loi [6] :***
Soit $(Y_n)_{n\geqslant 0}$ *une suite des variables aléatoires de fonction de répartition* F_{Y_n} *et* Y *une variable aléatoire de fonction de répartition* F_Y*. On dit que la suite* $(Y_n)_{n\geqslant 0}$ ***converge en loi*** *vers* Y *si pour tout* $t \in \mathbf{R}$ *tel que* F_Y *est continue en* t *:*

$$F_{Y_n}(t) \underset{n\to+\infty}{\longrightarrow} F_Y(t)$$

On la note :

$$Y_n \underset{n\to+\infty}{\xrightarrow{loi}} Y$$

1.4 Théorème central limite

L'un des outils les plus pratiques pour la construction d'un intervalle de confiance est le célèbre **théorème central limite**. Ce théorème est basé sur la convergence en loi que nous avons précédemment définie. C'est un théorème assez robuste car il permet de construire une variable aléatoire qui, sous certaines conditions souvent faciles à vérifier, converge vers une loi normale centrée réduite pour une taille d'échantillon n suffisamment grande ($\mathbf{n} \geq 30$).

Théoreme 1. ***Théorème central limite [6] :***
Soit $(Y_n)_{n\geq 0}$ *une suite des variables aléatoires indépendants et identiquement distribuées (iid) admettant un moment d'ordre deux. soit* σ^2 *la variance de* Y_1 *et* $\mathbf{E}(Y_1)$ *l'espérance de* Y_1*, alors :*

$$Y = \frac{\sqrt{n}}{\sqrt{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbf{E}(Y_1)\right) \underset{n\to+\infty}{\xrightarrow{loi}} Z \sim \mathcal{N}(0,\,1)$$

On peut donc déduire le corollaire, qui va nous conduire exactement à un intervalle de confiance, suivant :

Corollaire 1. *Soit* $(Y_n)_{n\geq 0}$ *une suite des variables aléatoires indépendants et identiquement distribuées (iid) admettant un moment d'ordre deux. soit* σ^2 *la variance de* Y_1 *et* $\mathbf{E}(Y_1)$ *l'espérance de* Y_1*, pour tout* $a > 0$*, on a :*

$$\mathcal{P}\left[-a < \frac{\sqrt{n}}{\sqrt{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbf{E}(Y_1)\right) < a\right] \underset{n\to+\infty}{\longrightarrow} \mathcal{P}[-a < Z < a] = F(a) - F(-a) = 2F(a) - 1$$

où F *est la distribution de la loi normale centré réduite.*

Il arrive souvent que la variance des variables aléatoires soit inconnue. Dans de telles situations, on peut appliquer une variante du théorème central limite, appelée **théorème central limite auto-normalisé**. Cette variante est pratiquement identique au théorème central, sauf qu'elle utilise un estimateur consistant de la variance à la place de la variance elle-même.

Théoreme 2. ***Théorème central limite auto-normalisé [6] :***
Soit $(Y_n)_{n\geq 0}$ *une suite des variables aléatoires indépendants et identiquement distribuées (iid) admettant un moment d'ordre deux. soit* $\hat{\sigma}^2$ *un* ***estimateur consistant*** *de la variance de* Y_1 *et* $\mathbf{E}(Y_1)$ *l'espérance de* Y_1*, alors :*

$$Y = \frac{\sqrt{n}}{\sqrt{\hat{\sigma}^2}} \left(\frac{1}{n} \sum_{i=1}^{n} Y_i - \mathbf{E}(Y_1) \right) \xrightarrow[n\to+\infty]{loi} Z \sim \mathcal{N}(0,\, 1)$$

Chapitre 2

Méthodes de calcul d'intervalles de confiance pour une proportion

2.1 Jeu des données

On considère $(X_i)_{1\leq i}$ une suite des variables aléatoires indépendantes et identiquement distribuées de la **loi bernouilli** de paramètre p succès, la suite $(X_i)_{1\leq i}$ est la population à étudier. On considère $(X_1, X_2, \cdots, X_n)$ des variables aléatoires un échantillon de cette population, de réalisation $(x_1, x_2, \cdots, x_n)$ et

$$t = \sum_{i=1}^{n} x_i$$

la somme d'observation dans cet échantillon.

2.2 Intervalle de confiance

2.2.1 Méthode standard de WALD

Cette méthode repose sur une approche normale, qui utilise la moyenne et l'écart-type de l'échantillon, ainsi que la distribution normale pour estimer la proportion de la population. Autrement dit, l'**intervalle de confiance de WALD** est calculé par inversion de la famille de régions de rejet du test statistique basé sur l'approximation de la loi binomiale par la loi normale centrée réduite. Cette méthode est particulièrement utile dans les cas où la taille de l'échantillon est suffisamment grande pour que l'approximation normale soit raisonnablement précise.

Hypothèses

Soit $p_0 \in [0, 1]$ une valeur donnée. On se demande si la proportion réelle, sur la population, p succès égale à la valeur fixé p_0. On pose l'hypothèse nulle $\mathcal{H}_0$ et l'hypothèse alternative $\mathcal{H}_1$ comme suit :

$$\mathcal{H}_0 : \quad p = p_0 \quad \text{contre} \quad \mathcal{H}_1 : \quad p \neq p_0$$

Statistique de test

Définition 7. ***Statistique de test :***
Une ***statistique de test*** *est une fonction des variable aléatoires* $(X_1, X_2, \cdots, X_n)$ *qui ne dépend pas du paramètre inconnu* p*.*

Dans notre cas , sous l'hypothèse $\mathcal{H}_0$ on pose la statistique de test T définie par :

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

où $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la moyenne empirique de l'échantillon et n la taille de l'échantillon.

Loi de la statistique de test

Les variables aléatoires $(X_1, X_2, \cdots, X_n)$ sont iid admettant un **espérance** $\mathbf{E}(X_1) = p$ paramètre de **bernouilli** et de variance $\mathcal{V}(X_1) = p(1-p)$ donc elles admettent un moment d'ordre deux. D'après la **loi forte des grands nombre**, la moyenne empirique de l'échantillon est un estimateur consistant de p. On considère la fonction continue dans $\mathbf{R}$ suivante :

$$\begin{aligned} \mathbf{f} : \mathbf{R} &\longrightarrow \mathbf{R} \\ x &\longmapsto \mathbf{f}(x) = x(1-x) \end{aligned}$$

Alors comme $\hat{p}$ est un estimateur consistant de p, sous la continuité de $\mathbf{f}$ dans $\mathcal{R}$, on a :

$$\hat{\mathcal{V}}(X_1) = \hat{p}(1-\hat{p})$$

un estimateur consistant de la variance de X_1.
D'après le **théorème central limite auto-normalisé**, on a :

$$T = \frac{\sqrt{n}(\bar{X_n} - p)}{\sqrt{\hat{\mathcal{V}}(X_1)}} \underset{n\to+\infty}{\overset{loi}{\sim}} \mathcal{N}(0,1)$$

En remplaçant la moyenne empirique $\bar{X_n}$ par son expression $\hat{p}$ et l'expression $\hat{\mathcal{V}}(X_1) = \hat{p}(1-\hat{p})$, on a :

$$T = \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}} \underset{n\to+\infty}{\overset{loi}{\sim}} \mathcal{N}(0,1)$$

Sous l'hypothèse nulle H_0, pour $p = p_0$, on a la statistique de test T qui suit la **loi normale centré réduite** :

$$T = \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} \underset{n\to+\infty}{\overset{loi}{\sim}} \mathcal{N}(0,1)$$

Sous l'hypothèse alternative H_1, pour $p \neq p_0$ et en s'appuyant sur la symétrie de la loi normale, la quantité $|T|$ prend des valeurs des plus en plus grandes (positives).

Nous sommes donc en mesure de construire la zone du rejet suivante :

Zone du rejet de niveau α

On a trouvé que la statistique de test suit la loi normale centré réduite sous l'hypothèse H_0. Sous l'hypothèse alternative H_1 la quantité $|T|$ est grande.

La zone du rejet est alors donnée par :

$$\mathcal{R} = \left\{ |T| > k_{1-\frac{\alpha}{2}} \right\} = \left\{ \left| \frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{\hat{p}(1-\hat{p})}} \right| > k_{1-\frac{\alpha}{2}} \right|\Bigg\}$$

On obtient la zone du rejet suivante :

$$\mathcal{R} = \left\{ |\hat{p} - p_0| > k_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right\}$$

où $k_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centré réduite.

Décision de WALD

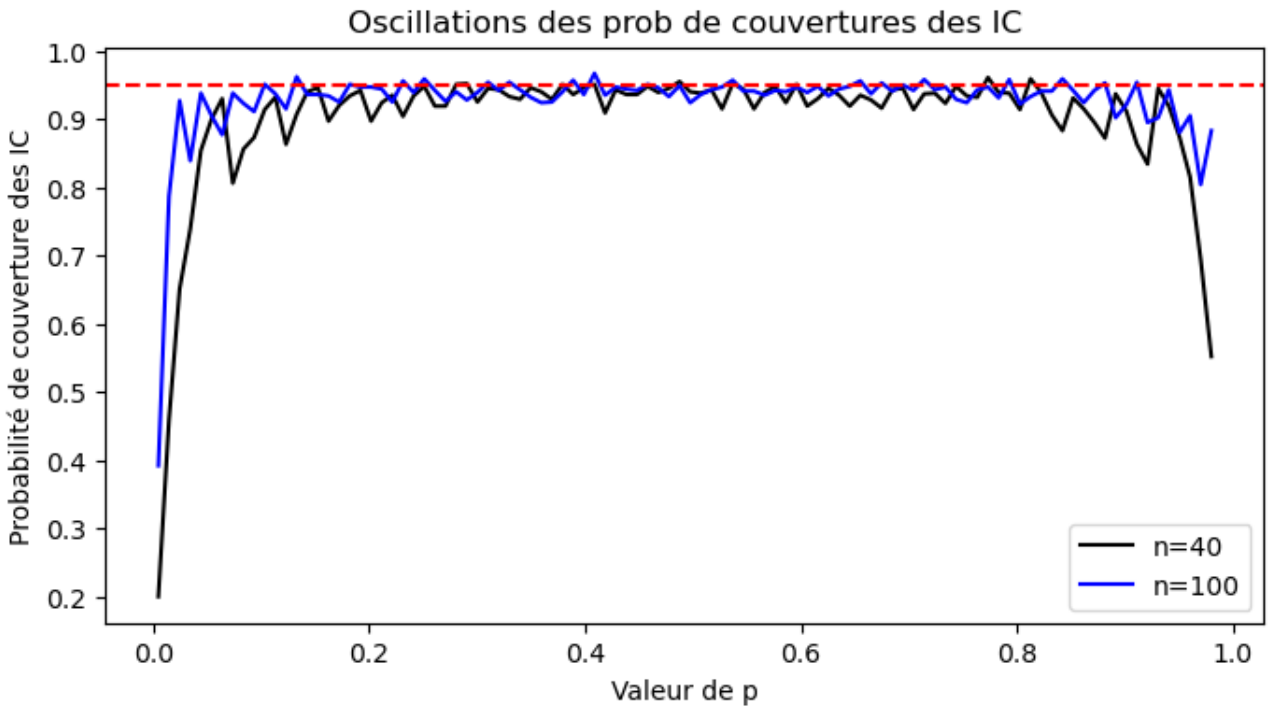
On accepte l'hypothèse alternative H_1 pour toute valeurs de $\hat{p}$ vérifiant l'inégalité de la zone $\mathcal{R}$.
On accepte l'hypothèse nulle H_0 pour toute valeur $\hat{p}$ qui ne vérifie pas l'inégalité de la zone $\mathcal{R}$. Ce dernière cas est équivalent à accepter l'hypothèse nulle H_0 pour toute valeur $\hat{p}$ telle que :

$$|\hat{p} - p_0| \leq k_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \Longrightarrow \quad \hat{p} - k_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p_0 \leq \hat{p} + k_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Toujours sous l'hypothèse H_0, la quantité p_0 est égale la quantité p proportion réelle de la population qu'on recherche. Alors l'intervalle de WALD s'écrit :

$$p = p_0 \in \left[\hat{p} - k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

où k la quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centré réduite.



Voir chap.3 partie 3.2.2) 1)

2.2.2 Méthode de Wald avec correction de continuité

la méthode de Wald pour l'intervalle de confiance d'une proportion peut produire des intervalles parfaitement centrés autour de la proportion empirique dans la plupart des cas. Cependant, pour des valeurs de p proches de 0 ou 1, la méthode peut produire une probabilité réelle quasiment inférieure au niveau nominal de confiance, comme le montre la figure ci-dessus. Ceci est dû à des biais potentiels qui peuvent survenir lorsque l'échantillon est de petite taille ou lorsque l'approximation de la distribution binomiale par la distribution normale centrée réduite est très faible. Dans ces cas, il est nécessaire de tronquer l'intervalle pour supprimer les valeurs aberrantes. L'approximation donnée par la méthode de Wald est précise pour des valeurs de p comprises entre 0,2 et 0,8.
En ce qui concerne la correction de l'intervalle, on ajoute $\frac{1}{2n}$ de chaque côté de l'intervalle pour corriger son caractère anti-conservateur. Cela permet de mieux couvrir la vraie proportion de la population lorsque la taille de l'échantillon est petite.

Ainsi pour l'**intervalle de confiance avec la correction de continuité**, on a :

$$p \in \left[\hat{p} - k\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} - \tfrac{1}{2n}, \hat{p} + k\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} + \tfrac{1}{2n}\right]$$

2.2.3 Méthode standard de score

L'intervalle de confiance de WALD n'est pas satisfaisant car la probabilité de recouvrement de la proportion réelle p erratique ; en particulier, outre de fortes oscillations liées au caractère discret de la loi binomiale lorsque $p < 0.2$ ou $p > 0.8$, il présente une probabilité réelle de recouvrement généralement bien inférieure du niveau nominal.
Selon WILSON, cet effet est du au centrage de l'intervalle en $\hat{p}$ qui n'est pas une bonne idée.

L'idée de WILSON est donc de recentrer l'intervalle fondé sur T_n, statistique de test de WALD, par son espérance approchée à l'ordre $o(1/n)$. Ceci dit qu'il remplace l'écart type estimé par WALD par l'écart type sous $\mathcal{H}_0$. On reprend les même étapes de la méthode WALD, notamment :

Hypothèses

$$\mathcal{H}_0 : \quad p = p_0 \quad \text{contre} \quad \mathcal{H}_1 : \quad p \neq p_0$$

Statistique de test

Comme méthode approchée, il se distingue de la méthode standard de WALD en remplaçant l'écart type estimé (Wald) par l'écart type sous $\mathcal{H}_0$: Sous l'hypothèse nulle $\mathcal{H}_0$, l'**écart type** est donc estimé par :

$$\sigma = p_0(1 - p_0)$$

Ainsi, sous l'hypothèse nulle $\mathcal{H}_0$, la statistique de WILSON est l'expression définie par :

$$Wl = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

où $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la moyenne empirique de l'échantillon et n la taille de l'échantillon.

Loi de la statistique de test

De même que la méthode de WALD standard, le **théorème central limite auto-normalisé** donne :

Sous l'hypothèse nulle H_0, pour $p = p_0$, on a la statistique de test Wl qui suit la **loi normale centré réduite** :

$$Wl = \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)}} \underset{n\to+\infty}{\overset{loi}{\sim}} \mathcal{N}(0, 1)$$

Sous l'hypothèse alternative H_1, pour $p \neq p_0$, la quantité $|Wl|$ prend des valeurs des plus en plus grandes (positive).

Zone du rejet de niveau α

De même que la zone du rejet de la méthode WALD, on obtient la zone du rejet :

La zone du rejet est alors donnée par :

$$\mathcal{R} = \{|Wl| > k\} = \{|\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}| > k|\}$$

On obtient la zone du rejet suivante :

$$\mathcal{R} = \left\{|\hat{p} - p_0| > k\sqrt{\tfrac{p_0(1-p_0)}{n}}\right\}$$

où k le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centré réduite.

Décision

On accepte l'hypothèse nulle H_0 pour toute valeur p_0 qui ne vérifie pas l'inégalité de la zone $\mathcal{R}$. Ce dernière est équivalent à accepter l'hypothèse nulle H_0 pour toute valeur p_0 telle que :

$$|\hat{p} - p_0| \leq k\sqrt{\frac{p_0(1-p_0)}{n}}$$

En élevant au carré cette inégalité, on trouve :

$$(\hat{p} - p_0)^2 - \frac{k}{n}p_0(1-p_0) \leq 0 \Longrightarrow (1 + \frac{k^2}{n})p_0^2 - (2\hat{p} + \frac{k^2}{n})p_0 + \hat{p}^2 \leq 0$$

Considérons l'équation :

$$(1 + \frac{k^2}{n})p_0^2 - (2\hat{p} + \frac{k^2}{n})p_0 + \hat{p}^2 = 0,$$

le déterminant est donné par :

$$\begin{aligned}\Delta &= (-2\hat{p} - \frac{k^2}{n})^2 - 4(1 + \frac{k^2}{n})\hat{p}^2 \\ &= 4\hat{p}^2 + \frac{k^4}{n} + 4\hat{p}\frac{k^2}{n} - 4\hat{p}^2 - 4\hat{p}^2\frac{k^2}{n} \\ &= \frac{k^4}{n^2} + 4\hat{p}(1-\hat{p})\frac{k^2}{n} \geqslant 0 \quad \text{car} \quad \hat{p} \in [0,1] \quad \text{donc} \quad (1-\hat{p}) \geqslant 0.\end{aligned}$$

Alors il existe deux solution en p_0 :

$$p_0^{(1)} = \frac{(2\hat{p} + \frac{k^2}{n}) - \sqrt{\Delta}}{2(1 + \frac{k^2}{n})} \quad \text{et} \quad p_0^{(2)} = \frac{(2\hat{p} + \frac{k^2}{n}) + \sqrt{\Delta}}{2(1 + \frac{k^2}{n})}$$

Pour $p_0^{(1)}$, on a :

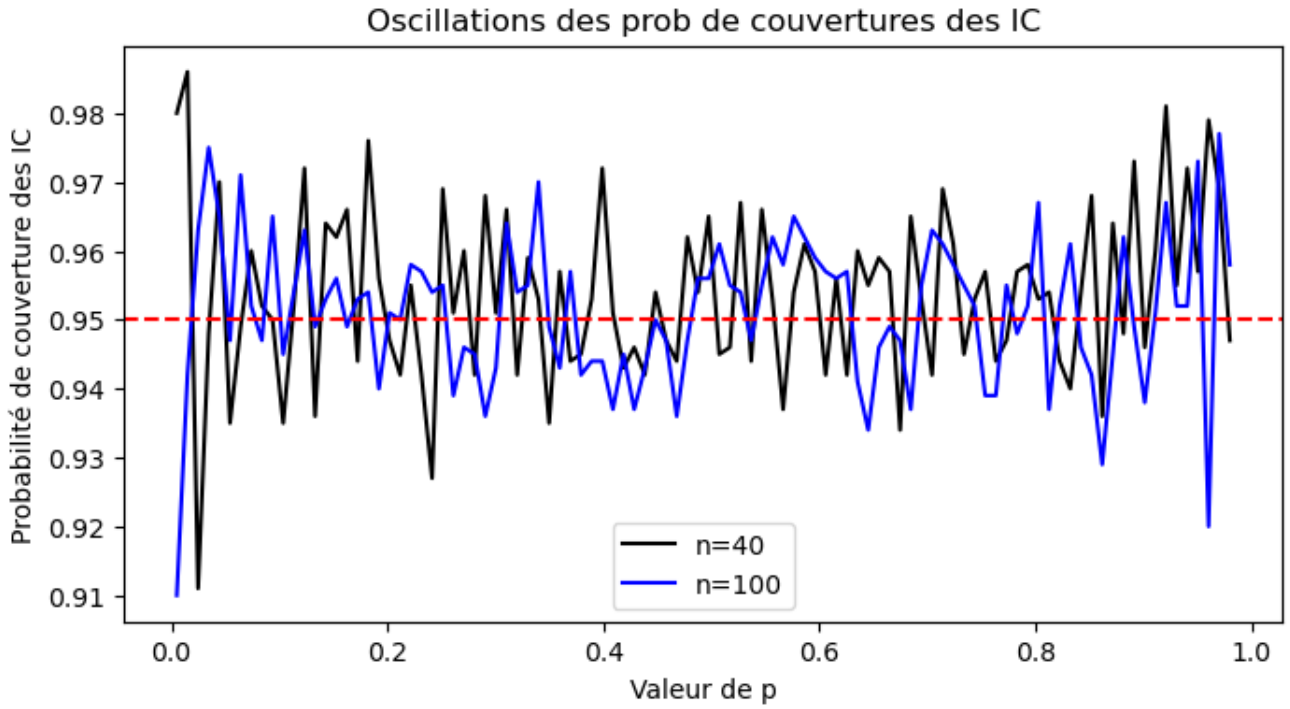
$$\begin{aligned} p_0^{(1)} &= \frac{(2\hat{p} + \frac{k^2}{n}) - \sqrt{\Delta}}{2(1 + \frac{k^2}{n})} \\ &= \frac{(2\hat{p} + \frac{k^2}{n}) - \sqrt{\frac{k^4}{n} + 4\hat{p}(1-\hat{p})\frac{k^2}{n}}}{2(1 + \frac{k^2}{n})} \\ &= \frac{(\hat{p} + \frac{k^2}{2n}) - \sqrt{\frac{k^4}{4n^2} + 4\hat{p}(1-\hat{p})\frac{k^2}{n}}}{1 + \frac{k^2}{2n}} \\ &= \frac{(\hat{p} + \frac{k^2}{2n}) - \frac{k}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p}) + \frac{k^2}{4n}}}{1 + \frac{k^2}{2n}} \end{aligned}$$

Alors de la même manière, on trouve $p_0^{(2)}$, définie par :

$$p_0^{(2)} = \frac{(\hat{p} + \frac{k^2}{2n}) + \frac{k}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p}) + \frac{k^2}{4n}}}{1 + \frac{k^2}{2n}}$$

L'hypothèse nulle H_0 suppose que $p = p_0$ alors l'**intervalle de confiance** pour la méthode de score aussi appelée méthode de **WILSON** est donnée par :

$$p = p_0 \in \left[\frac{(\hat{p}+\frac{k^2}{2n}) - \frac{k}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p})+\frac{k^2}{4n}}}{1+\frac{k^2}{2n}}, \frac{(\hat{p}+\frac{k^2}{2n}) + \frac{k}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p})+\frac{k^2}{4n}}}{1+\frac{k^2}{2n}} \right]$$



voir chap.3 partie 3.2.2) 2)

2.2.4 Méthode de score avec correction de continuité

De même que la méthode WALD standard, des effets de **biais** potentiel qui peuvent survenir à la petite taille de l'échantillon, ainsi qu'une très faible approximation de la distribution binomiale par la distribution normale centrée réduite se présentent. Pour résoudre ce cas échéant, WILSON propose une correction de continuité pour tenir compte du passage de la loi discrète à une loi continue. Chaque réalisation x sera considéré comme couvrant l'intervalle allant de $x-\frac{1}{2}$ à $x+\frac{1}{2}$. Cela induit une légère modification de la formule pour le calcul de l'intervalle de confiance, donné par :

Au niveau α l'expression de l'**intervalle du score avec correction de continuité** est donnée par :

$$p \in [\beta_-, \beta_+]$$

où

$$\beta_- = \frac{2n\hat{p} + k^2 - 1 - k\sqrt{k^2 - 2 - \frac{1}{n+4\hat{p}(n(1-\hat{p})+1)}}}{2(n+k^2)}$$

$$\beta_+ = \frac{2n\hat{p} + k^2 + 1 + k\sqrt{k^2 + 2 - \frac{1}{n+4\hat{p}(n(1-\hat{p})-1)}}}{2(n+k^2)}$$

ceci revient à dire :

$$p \in [\hat{\beta_-}, \hat{\beta_-}]$$

avec

$$\hat{\beta_-} = 2n\hat{p} + k^2 - 1 - k\sqrt{k^2 - 2 - \frac{1}{n+4\hat{p}(n(1-\hat{p})+1)}}$$

$$\hat{\beta_+} = 2n\hat{p} + k^2 + 1 + k\sqrt{k^2 + 2 - \frac{1}{n+4\hat{p}(n(1-\hat{p})-1)}}$$

2.2.5 Méthode d'Agresti-Coull

Théoriquement l'intervalle de confiance de WILSON donne un centrage correcte au tour de $\hat{p}$ puisque on trouve que :

$$E(Wl) = E\left(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}\right) = 0 \quad \text{pour tout} \quad n \in \mathcal{N}.$$

Cependant, l'expression de l'intervalle de confiance de WILSON est complexe et difficilement directement interprétable. Les deux auteurs, **Agresti** et **Coull**, ont observé que le centre de cet intervalle est une moyenne pondérée de l'estimateur $\hat{p}$ et $\frac{1}{2}$:

$$\tilde{P_{Ws}} = \hat{p}\left(\frac{n}{n+k^2}\right) + 0.5\left(\frac{k^2}{n+k^2}\right)$$

Cet estimateur tend à recentrer l'estimateur naturel vers 0.5 (rétrécissement).
Comme la valeur k est très proche de 2, $k = 1.96$, si l'on choisit une probabilité de recouvrement

$1-\alpha = 0.95$, **Agresti** et **Coull** ont proposé de rajouter artificiellement à l'échantillon 4 observations répondant à un tirage déterministe : la règle **+2 succès**, **+2 échecs**. On a alors le centrage de **Agresti** et **Coull** :

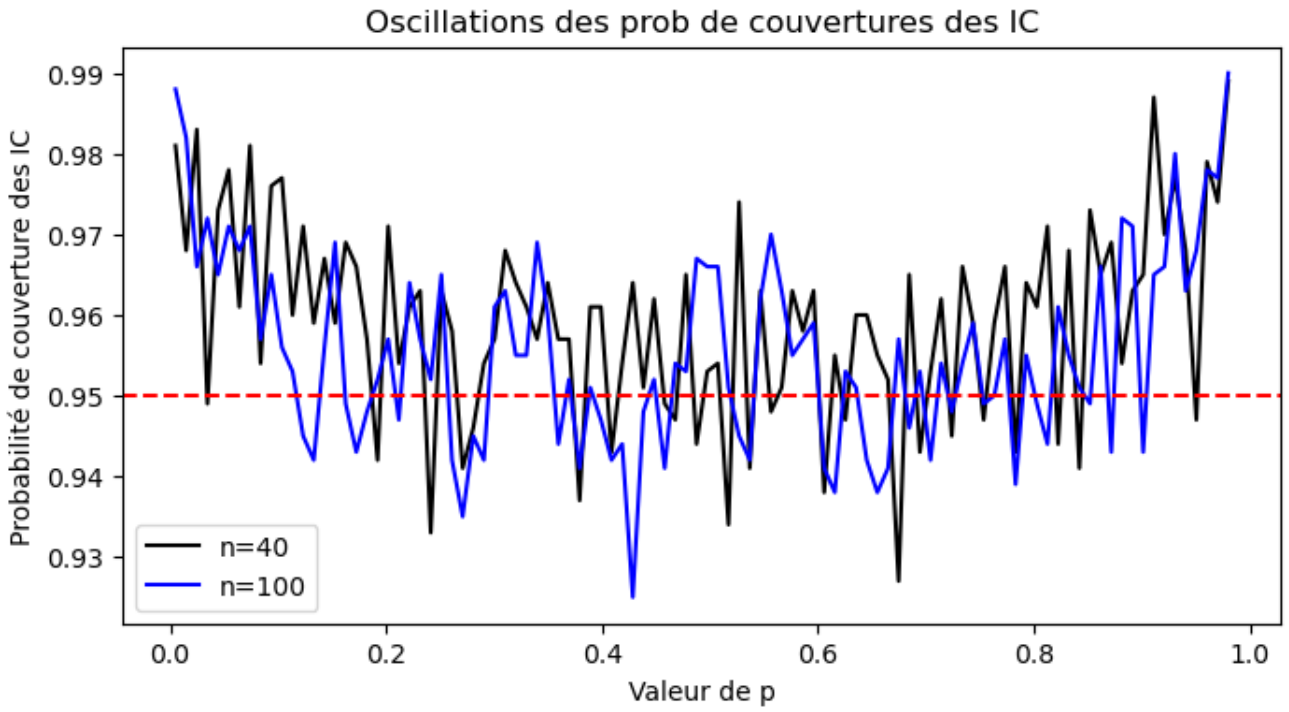
$$\tilde{P_{AC}} = \hat{p}\left(\frac{n}{n+4}\right) + 0.5\left(\frac{4}{n+4}\right)$$

En rajoutant les quatres variables aléatoires, de réalisation deux succès et deux échecs (de manière déterministe) à l'échantillon, l'intervalle d'**Agresti Coull** a une expression très simple :

$$p \in \left[\tilde{P_{AC}} + k\sqrt{\frac{\tilde{P_{AC}}(1-\tilde{P_{AC}})}{\tilde{n}}}, \tilde{P_{AC}} - k\sqrt{\frac{\tilde{P_{AC}}(1-\tilde{P_{AC}})}{\tilde{n}}}\right]$$

où $\tilde{n} = n+4$

Il est donc facile à présenter et très proche de l'intervalle de Wilson qui a un très faible biais de recouvrement (hors oscillations) en même temps qu'il est précis et que la probabilité de recouvrement réelle à tendance d'être au dessus du niveau nominal 0.95, comme on le montre la figure ci-dessous.



voir chap.3 partie 3.2.2) 3)

2.2.6 Méthode exacte de Clopper-Pearson

Les méthodes qui sont déjà annoncé sont toutes obtenues par le biais de l'approximation de la loi normale centrée réduite. Pour éviter d'avoir recours à une approximation, les ouvrages de statistique plus avancés, en particulier ceux orientés vers les applications industrielles recommandent l'intervalle « **exact** » de **Clopper-Pearson**. L'intervalle exacte de Clopper-Pearson

est basé, comme l'intervalle de Wilson, sur l'inversion de la famille des régions d'acceptation des tests,

$$\mathcal{H}_0 : \quad p = p_0 \quad \text{contre} \quad \mathcal{H}_1 : \quad p \neq p_0, \quad p_0 \quad \text{un réel dans [0,1]}$$

mais sans avoir recours à l'approximation normale de la loi binomiale.

L'idée est de borner le paramètre p dans un intervalle où les fluctuations de la probabilité réelle ne dépasse pas au dessous du niveau nominal de recouvrement.

Pour cela, en partant sous les hypothèses précédentes, on considère :

statistique de test

La statistique de test X_{CP} somme des variables aléatoire bernouilli :

$$X_{CP} = \sum_{i=1}^{n} X_i$$

une variable aléatoire de loi binomiale(n,p).

Loi de la statistique de test

Sous l'hypothèse $\mathcal{H}_0$, X_{CP} suit une binomiale de paramètre p_0.

Zone du rejet

Les deux auteurs **Clopper** et **Pearson** cherchent à mettre comme seuil, $l(p)$ et $u(p)$ appelés **limites de l'intervalle de fluctuations** de niveau $1-\alpha$. Comme on l'a dit tout avant, l'idée est d'avoir des fluctuations qui ne dépassent pas au dessous de la probabilité de recouvrement $1-\alpha$.
Ces deux seuils de Clopper-Pearson sont obtenues sous $\mathcal{H}_0$ par :

$$Pr_{p_0}(X_{CP} \leq l(p_0)) \leq \frac{\alpha}{2} \quad \text{et} \quad Pr_{p_0}(X_{CP} \geq u(p_0)) \leq \frac{\alpha}{2}$$

Les limites de confiance de Clopper-Pearson, sont obtenues en remplaçant les deux seuils par x :

$$x = \sum_{i=1}^{n} x_i$$

réalisation de la variable binomiale, on obtient :

$$Pr_{p_0}(X_{CP} \leq x) \leq \frac{\alpha}{2} \quad \text{et} \quad Pr_{p_0}(X_{CP} \geq x) \leq \frac{\alpha}{2}$$

Pour résoudre les deux inégalité, il est legitime de poser :

$$Pr_{p_0}(X_{CP} \leq x) = \frac{\alpha}{2} \quad \text{et} \quad Pr_{p_0}(X_{CP} \geq x) = \frac{\alpha}{2}$$

On sait que pour une variable aléatoire X binomiale, on a :

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \forall k = 0, 1, \cdots, n$$

Alors on en déduit :

$$\begin{aligned} Pr_{p_0}(X_{CP} \leq x) &= Pr_{p_0}(X_{CP} = 0) + Pr_{p_0}(X_{CP} = 1) + \cdots + Pr_{p_0}(X_{CP} = x) \\ &= \sum_{k=0}^{x} Pr_{p_0}(X_{CP} = k) \\ &= \sum_{k=0}^{x} \binom{n}{k} p_0^k (1-p_0)^{n-k} \end{aligned}$$

De même, on trouve :

$$\begin{aligned} Pr_{p_0}(X_{CP} \geq x) &= Pr_{p_0}(X_{CP} = x) + Pr_{p_0}(X_{CP} = x+1) + \cdots + Pr_{p_0}(X_{CP} = n) \\ &= \sum_{k=x}^{n} Pr_{p_0}(X_{CP} = k) \\ &= \sum_{k=x}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \end{aligned}$$

Les limites de confiance de Clopper-Pearson sont définies par les valeurs de p_0 solutions de :

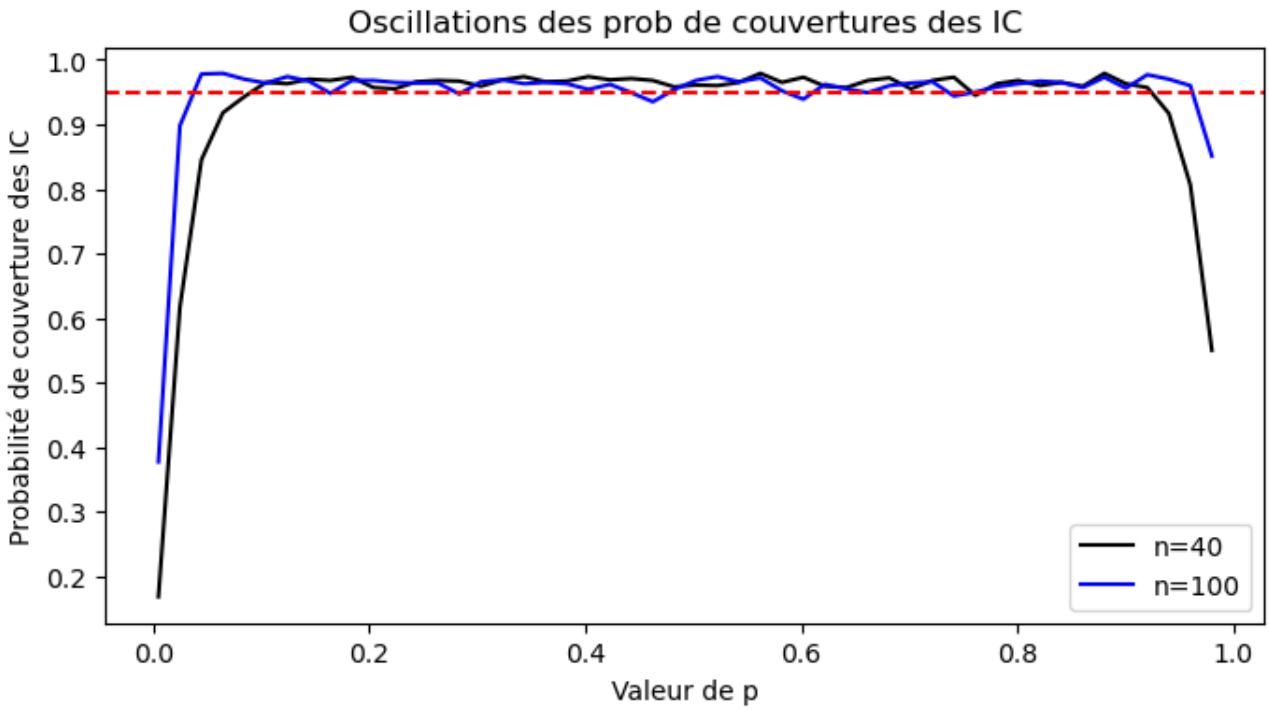
$$\sum_{k=0}^{x} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \frac{\alpha}{2} \quad \text{et} \quad \sum_{k=x}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \frac{\alpha}{2}$$

avec pour borne inférieure 0 lorsque $x = 0$ et pour borne supérieure 1 lorsque $x = 1$. Lorsque $x = 1, 2, ..., n - l$, l'intervalle de confiance est égal à :

$$p = p_0 \in [F_{\frac{\alpha}{2}, x, n-x+1}, F_{1-\frac{\alpha}{2}, x+1, n-x}]$$

où $F_{a,b,c}$ désigne le quantile c de la distribution **bêta** F avec des degrés de liberté a et b.

Comme on l'a dit avant, cet intervalle a la propriété de garantir une probabilité de recouvrement au moins égale à sa valeur nominale (voir figure ci-après).



voir chap.3 partie 3.2.3

Chapitre 3

Simulation

3.1 Méthodologie

Dans la suite, on considère un niveau de confiance $\alpha = 0.05$, cela dit qu'on travaille avec le quantile d'ordre 0.95 de la loi normal centrée réduite.

La fonction **graphe** a pour objectif de simuler la construction d'intervalles de confiance pour différentes valeurs de probabilité de succès p. Pour chaque valeur de p, la fonction génère un échantillon de taille n à partir d'une distribution de Bernoulli avec une probabilité de succès p donnée, puis calcule l'intervalle de confiance correspondant en utilisant la méthode de donnée.

Ensuite, la fonction vérifie si la vraie proportion p est couverte par l'intervalle de confiance et si c'est le cas, elle ajoute la probabilité de couverture correspondante à une liste. Ce processus est répété k fois pour chaque valeur de p, en générant toujours un nouvel échantillon de taille n à chaque fois. Cette procédure permet d'obtenir une liste des probabilités de couverture avec les valeurs de p correspondantes, ce qui fournit une estimation fiable de la probabilité de couverture de l'intervalle de confiance pour chaque valeur de p.

L'objectif principal de cette fonction est de comparer le niveau de couverture nominal de 0.95 fixé aux niveaux de couverture réels. Si une valeur de p appartient à un intervalle de confiance dont le niveau de couverture est inférieur à 0.95, cela signifie que la précision de l'intervalle de confiance associé à cette valeur de p n'est pas satisfaisante. En revanche, si une valeur de p appartient à un intervalle de confiance dont le niveau de couverture est supérieur à 0.95, cela indique que la précision de l'intervalle de confiance associé à cette valeur de p est satisfaisante.

3.2 Algorithmes

3.2.1 Bibliothèque

Des bibliothèque important pour la simulation sont :

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import beta
```

La première pour fonction mathématique comme racine carré, la deuxième pour l'affiche des graphe et la dernière pour le calcul du quantile de la loi bêta.

3.2.2 Méthode assymptotique

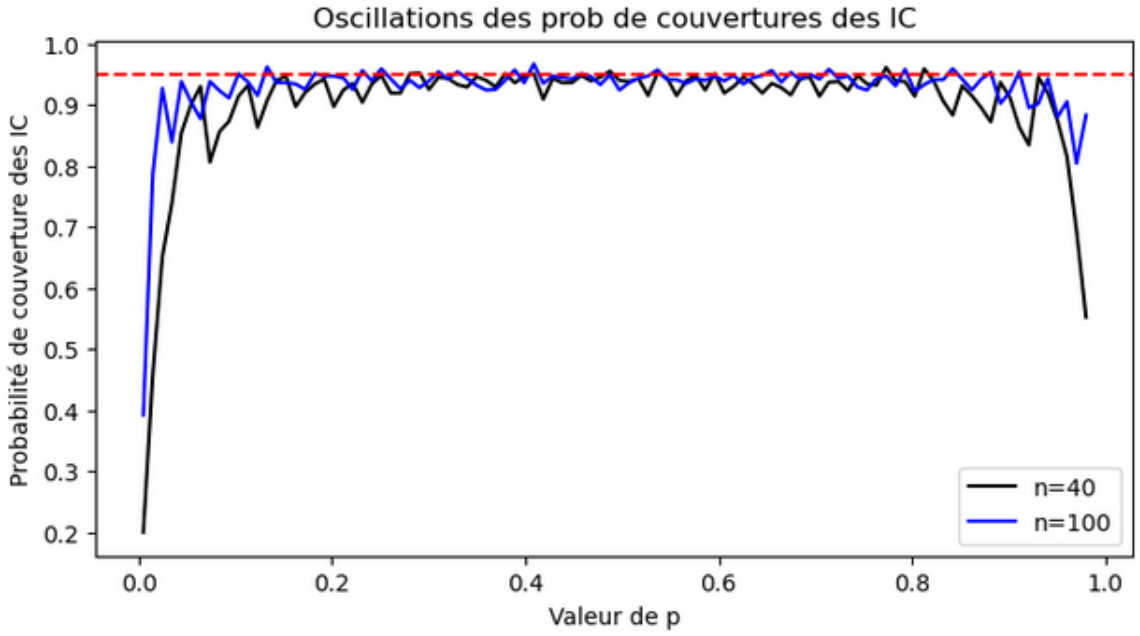
On utilise la fonction **graphe1** pour l'affichage de la méthode de **WALD**, **WILSON** et **Agresti-Coull** :

```
def graphe1(n1, n2, num_reps, methode):
    # Valeurs de p à considérer
    ps = np.linspace(0.005, 0.98, 100)
    # Probabilités de couverture pour chaque valeur de p
    proba_couverture1 = []
    proba_couverture2 = []
    for p in ps:
        i1 = 0
        i2 = 0
        for j in range(num_reps):
            # Générer un échantillon de taille n1 à partir d'une distribution Bernoulli
            sample1 = np.random.binomial(1, p, n1)
            # Calculer l'intervalle de confiance selon la méthode donnée
            born_inf1, born_sup1 = methode(np.mean(sample1), n1)
            # Vérifier si la vraie proportion p est couverte par l'intervalle de confian
            if p >= born_inf1 and p <= born_sup1:
                i1 += 1
            # Générer un échantillon de taille n2 à partir d'une distribution Bernoulli
            sample2 = np.random.binomial(1, p, n2)
            # Calculer l'intervalle de confiance selon la méthode donnée
            born_inf2, born_sup2 = methode(np.mean(sample2), n2)
            # Vérifier si la vraie proportion p est couverte par l'intervalle de confian
            if p >= born_inf2 and p <= born_sup2:
                i2 += 1
        # Ajouter la probabilité de couverture à la liste
        proba_couverture1.append(i1/num_reps)
        proba_couverture2.append(i2/num_reps)
    plt.figure(figsize=(8, 4))
    plt.plot(ps, proba_couverture1, color='black', label=f'n={n1}')
    plt.plot(ps, proba_couverture2, color='blue', label=f'n={n2}')
    plt.axhline(y=0.95, color='r', linestyle='--')
    plt.xlabel('Valeur de p')
    plt.ylabel('Probabilité de couverture des IC')
    plt.title('Oscillations des prob de couvertures des IC')
    plt.legend()
    plt.show()
```

1. Méthode de WALD

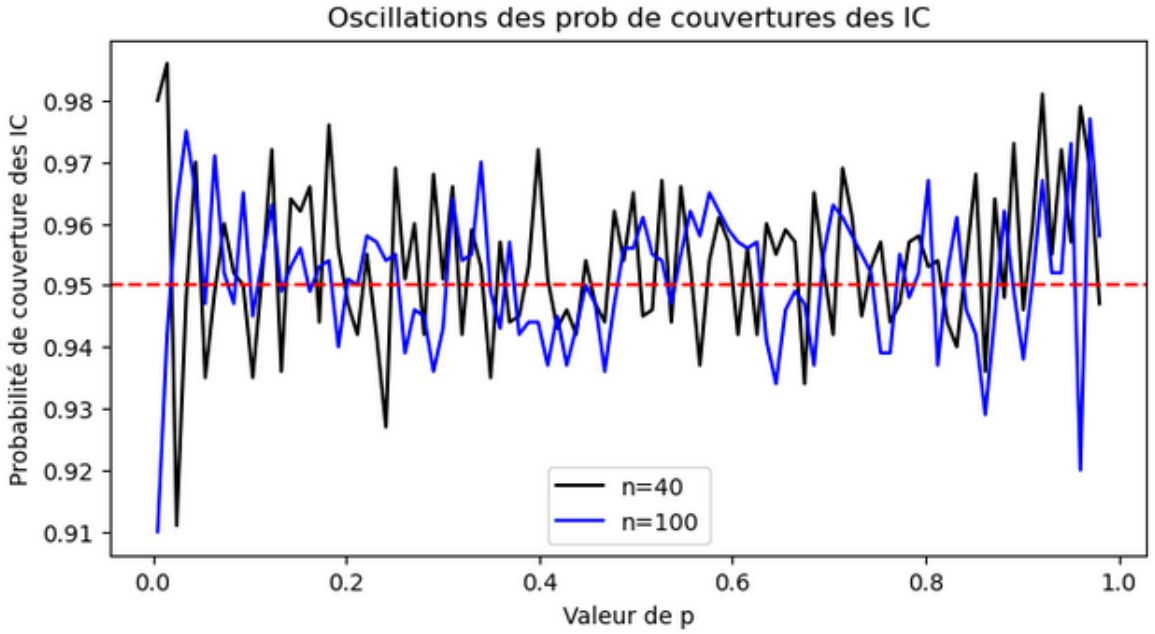
```
def wald_ci(p, n):
    z = 1.96
    born_inf = p - z*np.sqrt(p*(1-p)/n)
    born_sup = p + z*np.sqrt(p*(1-p)/n)
    return (born_inf, born_sup)
```

```
graphe1(40,100,1000,wald_ci)
```



2. Méthode de WILSON

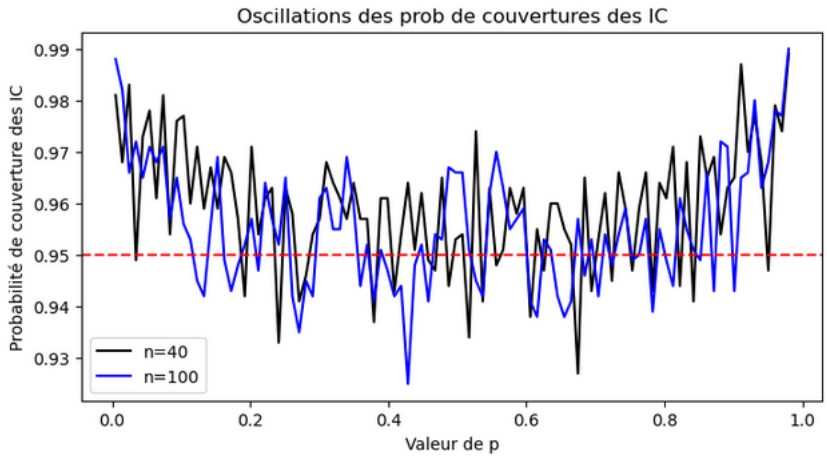
```
def wilson_ci(p, n):
    z = 1.96
    term = z*np.sqrt(p*(1-p)/n + z*z/(4*n*n))
    born_inf = (p + z*z/(2*n) - term)/(1+(z*z)/n)
    born_sup = (p + z*z/(2*n) + term)/(1+(z*z)/n)
    return born_inf, born_sup
```



3. Méthode de Agresti-Coull

```python
def agresti_coull_ci(p, n):

    # quantile de la loi normale standard pour le niveau de confiance alpha/2 avec
    z = 1.96
    A = p*(n/(n+4))+0.5*(4/(n+4))
    j = n + 4

    # Calculer l'écart-type estimé de la distribution de A
    B = z*np.sqrt(A*(1-A)/j)

    # Calculer l'intervalle de confiance
    born_inf = A - B
    born_sup = A + B

    return born_inf, born_sup
```

```python
graphe1(40,100,1000,agresti_coull_ci)
```



3.2.3 Méthode exacte

1. Méthode de Clopper-Pearson

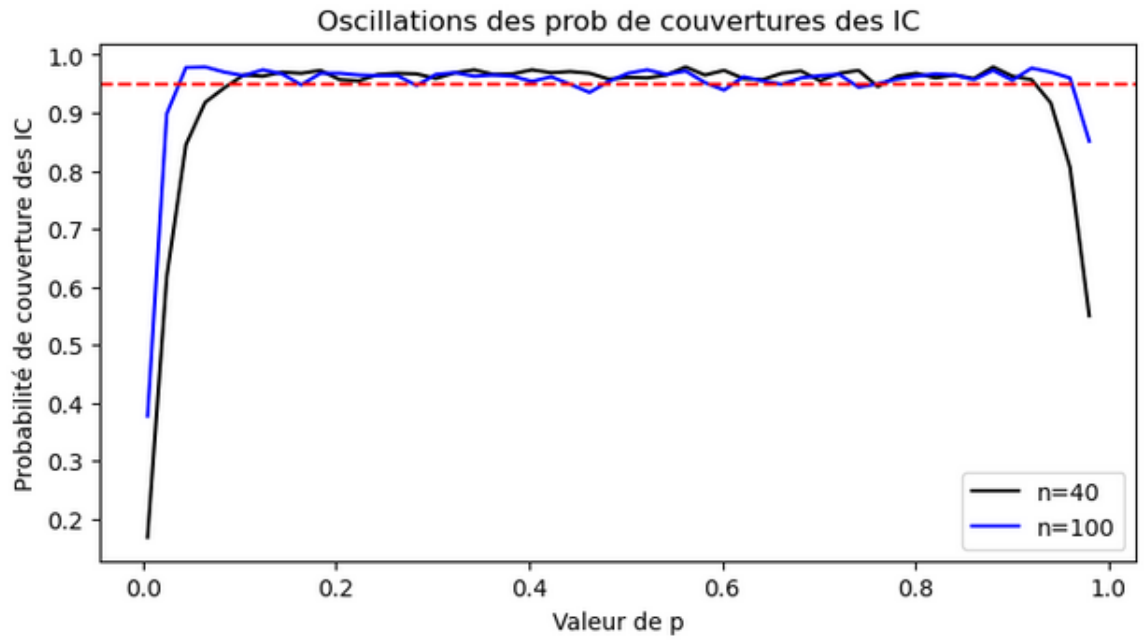
```python
  #  Calcul du quantiule de la loi beta
def quantile_beta(ordre, a, b):
    # degrés de liberté a et b

    # niveau de confiance ordre

    # calcul du quantile
    q = beta.ppf(ordre, a, b)

    return q
```

```python
def clopper_pearson(x, n):

    alpha = 0.05
    ordre_1 = alpha/2
    ordre_2 = 1-(alpha/2)

    born_inf = quantile_beta(ordre_1, x, (n-x+1))
    born_sup = quantile_beta(ordre_2, (x+1), (n-x))

    return born_inf, born_sup
```

On utilise la fonction **graphe2** pour l'affichage de la méthode de **Clopper-Pearson** :

```python
def graphe2(n1, n2, num_reps):
    ps = np.linspace(0.005, 0.98, 50)
    proba_couverture1 = []
    proba_couverture2 = []
    for p in ps:
        i1 = 0
        i2 = 0
        for j in range(num_reps):
            # Générer un échantillon de taille n1 à partir d'une distribution Bernoulli avec une proba
            sample1 = np.random.binomial(1, p, n1)
            # Nombre de succés
            somme1 = np.sum(sample1)
            # Calculer l'intervalle de confiance selon la méthode donnée
            born_inf1, born_sup1 = clopper_pearson(somme1, n1)
            # Vérifier si la vraie proportion p est couverte par l'intervalle de confiance
            if p >= born_inf1 and p <= born_sup1:
                i1 += 1
            # Générer un échantillon de taille n2 à partir d'une distribution Bernoulli avec une proba
            sample2 = np.random.binomial(1, p, n2)
            somme2 = np.sum(sample2)
            # Calculer l'intervalle de confiance selon la méthode donnée
            born_inf2, born_sup2 = clopper_pearson(somme2, n2)
            # Vérifier si la vraie proportion p est couverte par l'intervalle de confiance
            if p >= born_inf2 and p <= born_sup2:
                i2 += 1
        proba_couverture1.append(i1/num_reps)
        proba_couverture2.append(i2/num_reps)
    plt.figure(figsize=(8, 4))
    plt.plot(ps, proba_couverture1, color='black', label=f'n={n1}')
    plt.plot(ps, proba_couverture2, color='blue', label=f'n={n2}')
    plt.axhline(y=0.95, color='r', linestyle='--')
    plt.xlabel('Valeur de p')
    plt.ylabel('Probabilité de couverture des intervalles')
    plt.title('Probabilités de couverture des intervalles de confiance')
    plt.legend()
    plt.show()
```

```python
graphe2(40,100,1000)
```



Bibliographie

[1] AGRESTI, A. and COULL, *B.A. Approximate is better than "Exact" for interval estimation of Binomial proportions, American Statistician Vol. 52, No. 2 (mai 1998), pp.119–126.*

[2] CLOPPER C. J. et PEARSON E. S., *The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial, Biometrika, n°26(4), 1934, pp. 404-413.*

[3] WILSON E. B., *Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association n°22, 1927, pp. 209-212.*

[4] Lawrence D. Brown, T. Tony Cai and Anirban DasGupta, *Interval Estimation for a Binomial Proportion. Statistical Science 2001, Vol. 16, No. 2, 101–133.*

[5] Baba Thiam, *Statistique 2, cours dispensé en 2022-2023 université de Lille 1.*

[6] Rafael Butez, *Statistique 1 et Statistique décisionnelle cours dispensé en 2022-2023 université de Lille 1.*