Statistiques : estimation et tests

Cours de probabilités et statistiques

Master 1 MEEF - Mathématiques
Sorbonne Université

Séance (annulée) du 16 mars 2020

Que sont les statistiques ?

Les statistiques manipulent les mêmes outils mathématiques que les probabilités, avec un point de vue différent.

Démarche probabiliste

Caractéristiques du modèle connues. On chercher à prédire ce que l'on va observer.

Démarche statistique

On dispose d'observations, et on cherche à retrouver une information sur certaines caractéristiques inconnues du modèle.

Un exemple

Exemple : $(X_n)_{n\in\mathbf{N}}$ des variables aléatoires de Bernoulli indépendantes et de même paramètre p.

Démarche probabiliste

p est connu, et on cherche à décrire a priori la suite $(X_n)_{n\in\mathbf{N}}$.
Par exemple, on montre que, presque sûrement, on a $\frac{1}{N}\sum_{n=1}^{N} X_n \to p$.

Démarche statistique

On observe une réalisation de $X_1,\ldots, X_N$, mais maintenant p est inconnu, et on cherche à retrouver des informations dessus (que vaut-il ? a-t-on $p > 1/2$?). On dispose par exemple d'une approximation $p \simeq \frac{1}{N}\sum_{n=1}^{N} X_n$.

Deux questions statistiques

Deux types de questions peuvent se poser en statistiques :

Estimation

On cherche à retrouver une valeur approchée d'un paramètre inconnu.

Test

On cherche à répondre oui/non à une question portant sur les paramètres (ex : $p > 1/2$? $\mu > 0$? $\lambda > 1$?).

Plan

Estimation

On cherche à retrouver une valeur approchée d'un paramètre inconnu.
Important : on souhaite également connaître l'ordre de grandeur de l'erreur commise (dire "$p \simeq 0.3$" n'a pas vraiment de sens, en revanche "$p \simeq 0.3 \pm 0.01$" est plus pertinent).

Définition : intervalle de confiance

Un intervalle de confiance de niveau $1-\alpha$, pour un paramètre c est un intervalle $\hat{I}$ obtenu à partir des observations (donc aléatoire), tel que

$$\mathbf{P}(c \in \hat{I}) \geq 1 - \alpha.$$

Remarques

- Dans l'évènement $\{c \in \hat{I}\}$, c'est bien $\hat{I}$ qui est aléatoire. c a beau être inconnu, il est tout de même fixé.
- Un "intervalle aléatoire" est simplement un intervalle de la forme $[X, Y]$, où X et Y sont des variables aléatoires : $\mathbf{P}(X \leq c \leq Y) \geq 1 - \alpha$.

Estimation

On cherche à retrouver une valeur approchée d'un paramètre inconnu.
Important : on souhaite également connaître l'ordre de grandeur de l'erreur commise (dire "$p \simeq 0.3$" n'a pas vraiment de sens, en revanche "$p \simeq 0.3 \pm 0.01$" est plus pertinent).

Définition : intervalle de confiance

Un intervalle de confiance de niveau $1-\alpha$, pour un paramètre c est un intervalle $\hat{I}$ obtenu à partir des observations (donc aléatoire), tel que

$$\mathbf{P}(c \in \hat{I}) \geq 1 - \alpha.$$

Remarques

- $\hat{I}$: valeur approchée avec marge d'erreur,
  α : probabilité de faire une prédiction incorrecte.
- En pratique on choisit α petit (certitude), et on voudrait l'intervalle $\hat{I}$ soit étroit (précision). Ces deux demandes sont antagonistes ! On doit faire un compromis.

Exemples de constructions d'intervalles de confiance : Bienaymé-Tchebychev

Soit $(X_n)_{n\in\mathbf{N}}$ une suite de variables aléatoires indépendantes, de carré intégrable, de même loi, de moyenne μ inconnue, et de variance σ^2 connue (exemple : quantité mesurée par un outil dont on connaît la précision).

Par l'inégalité de Bienaymé-Tchebychev, on a :

$$\mathbf{P}\left(\left|\mu - \frac{1}{N}\sum_{n=1}^{N} X_n\right| \leq \quad \varepsilon \quad\right) \geq 1 - \frac{\sigma^2}{N\varepsilon^2}.$$

Exemples de constructions d'intervalles de confiance : Bienaymé-Tchebychev

Soit $(X_n)_{n\in\mathbf{N}}$ une suite de variables aléatoires indépendantes, de carré intégrable, de même loi, de moyenne μ inconnue, et de variance σ^2 connue (exemple : quantité mesurée par un outil dont on connaît la précision).

Par l'inégalité de Bienaymé-Tchebychev, on a :

$$\mathbf{P}\left(\left|\mu - \frac{1}{N}\sum_{n=1}^{N} X_n\right| \leq \frac{\sigma}{\sqrt{\alpha N}}\right) \geq 1 - \alpha \ .$$

Exemples de constructions d'intervalles de confiance : Bienaymé-Tchebychev

Soit $(X_n)_{n\in\mathbf{N}}$ une suite de variables aléatoires indépendantes, de carré intégrable, de même loi, de moyenne μ inconnue, et de variance σ^2 connue (exemple : quantité mesurée par un outil dont on connaît la précision).

Par l'inégalité de Bienaymé-Tchebychev, on a :

$$\mathbf{P}\left(\left|\mu - \frac{1}{N}\sum_{n=1}^{N} X_n\right| \leq \frac{\sigma}{\sqrt{\alpha N}}\right) \geq 1 - \alpha\ .$$

Ceci peut se récrire :

$$\mathbf{P}\left(\mu \in \left[\frac{1}{N}\sum_{n=1}^{N} X_n - \frac{\sigma}{\sqrt{\alpha N}}, \frac{1}{N}\sum_{n=1}^{N} X_n + \frac{\sigma}{\sqrt{\alpha N}}\right]\right) \geq 1 - \alpha.$$

Autrement dit, l'intervalle en vert est un intervalle de confiance de niveau $1-\alpha$ pour μ.

Exemples de constructions d'intervalles de confiance : variables Gaussiennes

Soit $(X_n)_{n\in\mathbf{N}}$ une suite de variables aléatoires indépendantes, Gaussiennes de moyenne μ inconnue, et de variance σ^2 connue.

Comme une somme de Gaussiennes indépendantes est Gaussienne, on a :

$$\mathbf{P}\left(\left|\mu - \frac{1}{N}\sum_{n=1}^{N} X_n\right| \leq \frac{c_\alpha \sigma}{\sqrt{N}}\right) = \int_{-c_\alpha}^{c_\alpha} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = 1-\alpha.$$

Ceci peut se récrire :

$$\mathbf{P}\left(\mu \in \left[\frac{1}{N}\sum_{n=1}^{N} X_n - \frac{c_\alpha \sigma}{\sqrt{N}}, \frac{1}{N}\sum_{n=1}^{N} X_n + \frac{c_\alpha \sigma}{\sqrt{N}}\right]\right) = 1-\alpha.$$

Autrement dit, l'intervalle en vert est un intervalle de confiance de niveau $1-\alpha$ pour μ.

Exemples de constructions d'intervalles de confiance : le théorème limite central

Soit $(X_n)_{n\in\mathbf{N}}$ une suite de variables aléatoires indépendantes, de carré intégrable, de moyenne μ inconnue, et de variance σ^2 connue.

Par le théorème limite central, on a :

$$\lim_N \mathbf{P}\left(\left|\mu - \frac{1}{N}\sum_{n=1}^{N} X_n\right| \leq \frac{c_\alpha \sigma}{\sqrt{N}}\right) = \int_{-c_\alpha}^{c_\alpha} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = 1-\alpha.$$

Ceci peut se récrire :

$$\lim_N \mathbf{P}\left(\mu \in \left[\frac{1}{N}\sum_{n=1}^{N} X_n - \frac{c_\alpha \sigma}{\sqrt{N}}, \frac{1}{N}\sum_{n=1}^{N} X_n + \frac{c_\alpha \sigma}{\sqrt{N}}\right]\right) = 1-\alpha.$$

Autrement dit, l'intervalle en vert est un intervalle de confiance asymptotique de niveau $1-\alpha$ pour μ.

Exemples de constructions d'intervalles de confiance : lois de Bernoulli

Soit $(X_n)_{n\in\mathbf{N}}$ une suite de variables aléatoires de loi de Bernoulli, de paramètre p inconnu.

Par le théorème limite central, on a :

$$\lim_N \mathbf{P}\left(\left|p - \frac{1}{N}\sum_{n=1}^{N} X_n\right| \leq \frac{c_\alpha\sqrt{p(1-p)}}{\sqrt{N}}\right) = \int_{-c_\alpha}^{c_\alpha} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = 1-\alpha.$$

Ceci peut se récrire :

$$\lim_N \mathbf{P}\left(\mu \in \left[\frac{1}{N}\sum_{n=1}^{N} X_n - \frac{c_\alpha\sqrt{p(1-p)}}{\sqrt{N}}, \frac{1}{N}\sum_{n=1}^{N} X_n + \frac{c_\alpha\sqrt{p(1-p)}}{\sqrt{N}}\right]\right) = 1-\alpha$$

Mais l'intervalle en vert n'est pas connu (il dépend de p).

Exemples de constructions d'intervalles de confiance : lois de Bernoulli

Soit $(X_n)_{n\in\mathbf{N}}$ une suite de variables aléatoires de loi de Bernoulli, de paramètre p inconnu.

Par le théorème limite central, on a :

$$\lim_N \mathbf{P}\left(\left|p - \frac{1}{N}\sum_{n=1}^{N} X_n\right| \leq \frac{c_\alpha\sqrt{p(1-p)}}{\sqrt{N}}\right) = \int_{-c_\alpha}^{c_\alpha} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = 1-\alpha.$$

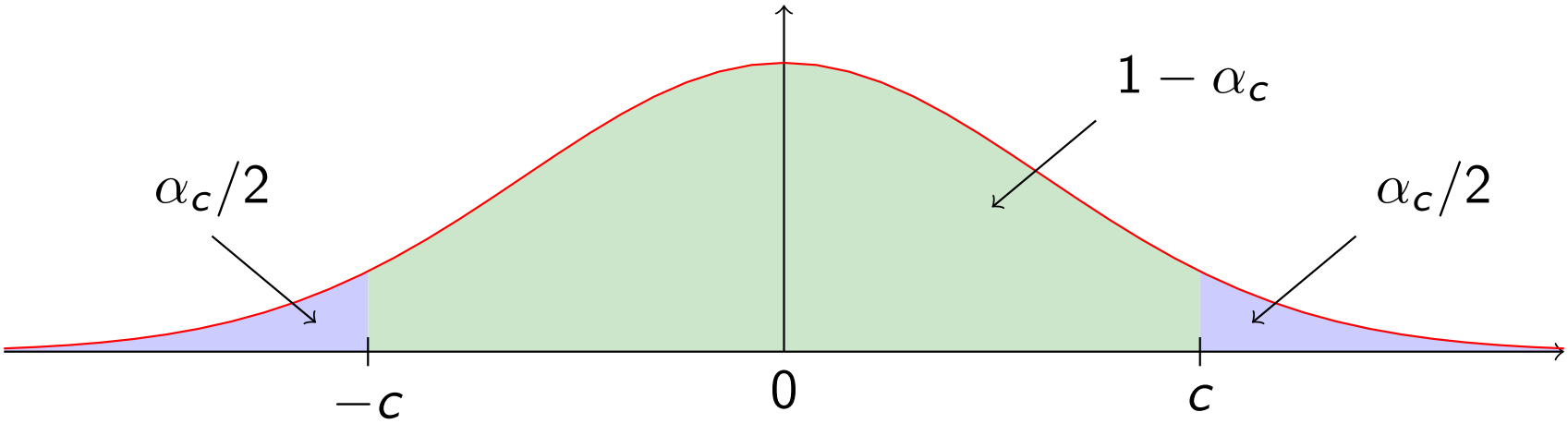
Ceci peut se récrire :

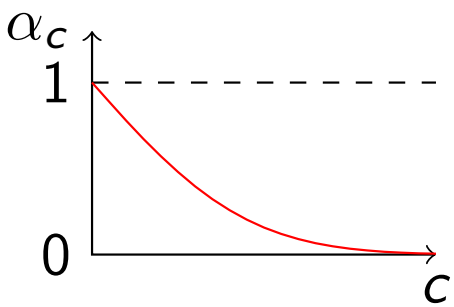
$$\lim_N \mathbf{P}\left(\mu \in \left[\frac{1}{N}\sum_{n=1}^{N} X_n - \frac{c_\alpha}{2\sqrt{N}}, \frac{1}{N}\sum_{n=1}^{N} X_n + \frac{c_\alpha}{2\sqrt{N}}\right]\right) \geq 1-\alpha.$$

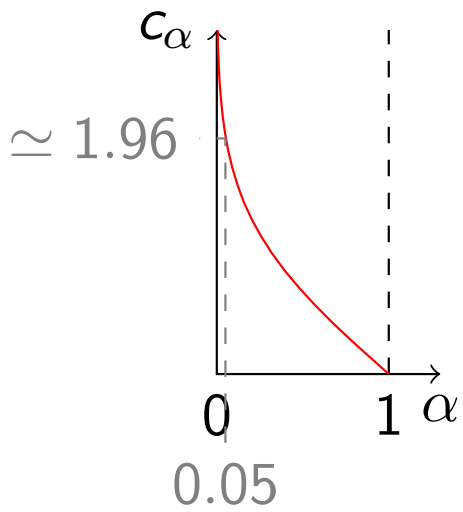
On utilise l'inégalité $\sqrt{p(1-p)} \leq 1/2$: l'intervalle obtenu est plus grand(=moins précis) que le précédent, et a donc une probabilité supérieure. On a bien obtenu un intervalle de confiance (asymptotique).

Valeur de c_α

Aire α_c sous la densité de loi normale centrée réduite $(e^{-x^2/2}(2\pi)^{-1/2})$, entre les points $-c$ et c :







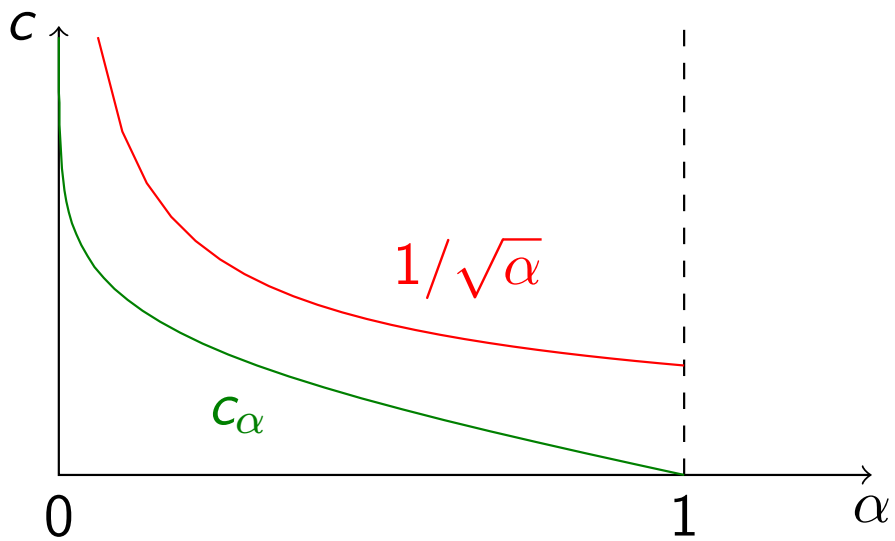
À gauche : la fonction $c \mapsto \alpha_c$, à droite, sa réciproque $\alpha \mapsto c_\alpha$.

Approximation Gaussienne vs. Bienaymé-Tchebychev

Largeur de l'intervalle de confiance à α fixé :

- Bienaymé-Tchebychev : $\frac{1}{\sqrt{\alpha}} \times \frac{\sigma}{\sqrt{N}}$,
- l'approximation Gaussienne : $c_\alpha \times \frac{\sigma}{\sqrt{N}}$ (beaucoup plus petit).

Autrement dit, à niveau fixé, l'approximation Gaussienne est bien plus précise que Bienaymé-Tchebychev :



Par exemple, pour $\alpha = 0.05$: $1/\sqrt{0.05} \simeq 4.47$, $c_{0.05} \simeq 1.96$.

Plan

Tests

Autre question statistique que l'on peut se poser : répondre à une question en oui/non, ou plus généralement, choisir entre deux possibilités.

Exemples

- $(X_n)_{n\in\mathbf{N}}$ indépendantes de loi de Bernoulli de paramètre p inconnu. A-t-on $p > 1/2$? (élection)
- $(X_n)_{n\in\mathbf{N}}$ des Gaussiennes indépendantes de moyenne $\mu \geq 0$ inconnue et de variance σ^2 connue. A-t-on $\mu > 0$ ou $\mu = 0$? (effet d'un médicament)
- $(X_n)_{n\in\mathbf{N}}$ des variables de même loi de moyenne inconnue $\mu \in \{\mu_1, \mu_2\}$ et de variance connue σ^2. A-t-on $\mu = \mu_1$ ou $\mu = \mu_2$? (identification d'une population)

Tests

Autre question statistique que l'on peut se poser : répondre à une question en oui/non, ou plus généralement, choisir entre deux possibilités.

Vocabulaire

Dans le vocabulaire statistique, plutôt que d'appeler les deux réponses "oui" ou "non", on parle plutôt d'hypothèse H_0 et d'hypothèse H_1.
Mettre en place un test, c'est définir une fonction qui a des observations associe une des valeurs H_0 ou H_1.

Exemples

- H_0 :"$p > 1/2$", H_1 :"$p \leq 1/2$",
- H_0 :"$\mu = 0$", H_1 :"$\mu > 0$",
- H_0 :"$\mu = \mu_1$", H_1 :"$\mu = \mu_2$",

Erreurs

Comme pour l'estimation, il y a un risque d'erreur, que l'on voudrait minimiser.

Deux types d'erreurs possibles : décider H_1, alors que c'est H_0 qui est vraie (erreur de première espèce) et inversement (erreur de deuxième espèce).

Si dans les cas incertains, on décide plus facilement H_0, on évite l'erreur de première espèce, mais on risque plus l'erreur de deuxième espèce. Si on répond plus facilement H_1, c'est l'inverse.
En particulier, on ne peut pas en même temps, minimiser les deux probabilités d'erreurs : il faut choisir d'éviter "à tout prix" une des deux erreurs, au détriment de l'autre.

Erreurs

En pratique, on fait en sorte de contrôler l'erreur de première espèce. Autrement dit, on dit qu'un test a un niveau $1-\alpha$ si lorsque H_0 est vraie, on décide H_0 avec probabilité $\geq 1-\alpha$.

En résumé

Décider H_1 : on est (presque) certain que l'hypothèse H_0 est fausse.
Décider H_0 : choix par défaut signifiant que les données ne sont pas suffisamment précises pour affirmer que H_0 est fausse : H_0 est crédible.

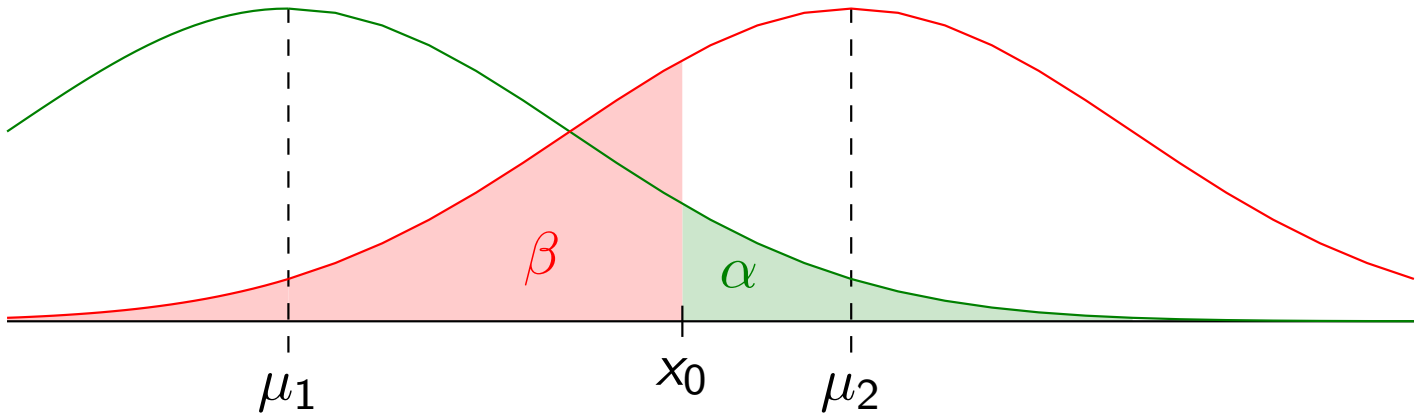
Un exemple

X suit la loi $\mathcal{N}(\mu, 1)$, avec $\mu \in \{\mu_1, \mu_2\}$, $\mu_1 < \mu_2$. Les deux hypothèses sont H_0 :"$\mu = \mu_1$", H_1 :"$\mu = \mu_2$".

En observant X, on veut tester les hypothèses H_0 et H_1. Comme $\mu_1 < \mu_2$ on va décider H_0 si $X < x_0$ pour une certaine valeur x_0 à déterminer, et H_1 si $X > x_0$.

Pour ne se tromper qu'avec probabilité α quand H_0 est vraie, on choisit x_0 tel que l'aire en vert sur le graphe ci-dessous vaille α.

Aire en rouge β : probabilité de commettre l'erreur de seconde espèce lorsque H_1 est vraie.



On dit que $]-\infty, x_0[$ est la région d'acceptation, et que $]x_0, \infty[$ est la région de rejet.

Autre exemple

$(X_n)_{n\in\mathbf{N}}$ une suite de variables de Bernoulli de paramètre p inconnu. On veut tester H_0 :"$p = 1/2$" contre H_1 :"$p \neq 1/2$". Si on est sous H_0 (c'est-à-dire si $p = 1/2$), le théorème limite central donne

$$\lim_N \mathbf{P}\left(\frac{1}{N}\sum_{n=1}^{N} X_n \in \left[\frac{1}{2} - \frac{c_\alpha}{2\sqrt{N}}, \frac{1}{2} + \frac{c_\alpha}{2\sqrt{N}}\right]\right) = 1 - \alpha.$$

On va donc décider H_0 si la moyenne empirique $\frac{1}{N}\sum_{n=1}^{N} X_n$ prend sa valeur dans l'intervalle $\left[\frac{1}{2} - \frac{c_\alpha}{2\sqrt{N}}, \frac{1}{2} + \frac{c_\alpha}{2\sqrt{N}}\right]$, et décider H_1 sinon.
Autrement dit, la région d'acceptation est $\left[\frac{1}{2} - \frac{c_\alpha}{2\sqrt{N}}, \frac{1}{2} + \frac{c_\alpha}{2\sqrt{N}}\right]$. Les précédents programmes de lycée donnaient le nom d'intervalle de fluctuation à cette région d'acceptation.
Par exemple, avec $N = 1000$, $\alpha = 5\%$, la région d'acceptation est $[0.469, 0.531]$.

Encoure un exemple

$(X_n)_{n\in\mathbf{N}}$ une suite de variables de Bernoulli de paramètre p inconnu. On veut tester H_0 :"$p < 1/2$" contre H_1 :"$p > 1/2$". Si on est sous H_0 (c'est-à-dire si $p < 1/2$), le théorème limite central donne

$$\lim_N \mathbf{P}\left(\frac{1}{N}\sum_{n=1}^{N} X_n > p + \frac{c_{2\alpha}\sqrt{p(1-p)}}{\sqrt{N}}\right) = \alpha.$$

Or

$$p + \frac{c_{2\alpha}\sqrt{p(1-p)}}{\sqrt{N}} < \frac{1}{2} + \frac{c_{2\alpha}}{2\sqrt{N}}.$$

On va donc décider H_0 si la moyenne empirique $\frac{1}{N}\sum_{n=1}^{N} X_n$ prend sa valeur dans l'intervalle $\left[0, \frac{1}{2} + \frac{c_{2\alpha}}{2\sqrt{N}}\right]$, et décider H_1 sinon. Autrement dit, la région d'acceptation est $\left[0, \frac{1}{2} + \frac{c_{2\alpha}}{2\sqrt{N}}\right]$.
Par exemple, avec $N = 1000$, $\alpha = 5\%$, la région d'acceptation est $[0, 0.526]$.