Comparaison de deux médicaments

Thomas BREDA et Mahendra MARIADASSOU
sous la direction de Patricia REYNAUD-BOURET

16 octobre 2004

Table des matières

1 Introduction **2**
1.1 Présentation du problème 2
1.2 Définitions et notations 2
1.2.1 Variables aléatoires de Bernoulli, Gaussienne 2
1.3 Retour au problème 3
1.3.1 Test d'hypothèses 3

2 Etude du premier modèle **4**
2.1 Niveau et puissance 6
2.2 Inversion de la figure 9

3 Etude du deuxième modèle **10**
3.1 Approximation Poissonnienne 10
3.2 Première approche 12
3.3 Ajustement du modèle 16
3.4 Niveau et puissance 18

Résumé

Cet exposé présente quelques méthodes statistiques utilisées dans le but de comparer deux médicaments. Le problème concret est le suivant : un échantillon de personnes âgées va être suivi médicalement pendant quelques mois à quelques années. Une partie sera traitée à l'aspirine, l'autre aux anticoagulants. Ces deux types de médicaments sont supposés prévenir les accidents cérébraux de type thrombose (un vaisseau du cerveau se bouche). La seule manière fiable de déceler un tel accident est un examen coûteux de type IRM. En revanche, les thromboses entraînent une perte des facultés mentales qui peut se mesurer à l'aide d'un simple questionnaire.

La question est donc : comment planifier cette expérience, avant qu'elle ait lieu, pour tester lequel des deux médicaments est le meilleur ?

Le problème sous-jacent est qu'il faut convaincre les médecins de faire subir à leurs patients cette expérience. La raison principale de leur réticence vient du fait suivant : en pratique ils se rendent assez vite compte que tel médicament marche mieux qu'un autre, et pourtant vu la planification initiale, ils sont obligés de donner toujours le même médicament au même patient pour ne pas fausser l'expérience. Il faudrait donc trouver une planification qui résolve aussi ce problème.

Ce problème médical est posé par le Dr Abastado (hôpital Pompidou).

1 Introduction

1.1 Présentation du problème

On se propose de traiter deux groupes distincts de patients avec deux médicaments différents et de décider, à partir des observations, lequel des deux est le plus efficace. On considère dans un premier temps un test simple : chaque patient est traité durant toute la durée de l'expérience avec un des deux médicaments et on relève à la fin de l'expérience le nombre de cas de thromboses pour chaque groupe à l'aide du mini-mental test (MMT). Ce test élémentaire permet d'évaluer de façon simple les facultés mentales du patient. On interprètera donc une chute au MMT comme la conséquence d'une thrombose.

Remarque 1.1.1. Il n'est pas évident qu'une chute au MMT soit due à une thrombose. En effet d'autres maladies liées au vieillissement (Alzheimer par exemple) peuvent altérer les facultés mentales. Cependant, le Dr Abastado semble dire que concrètement, le MMT est le seul test praticable à un coût raisonnable.

Formellement, on considère n_1 patients traités au médicament m_1 (groupe 1) et n_2 patients traités au médicament m_2 (groupe 2).
On considère les v.a $(X_i)_{i=1..n_1}$ et $(Y_j)_{j=1..n_2}$ où

$$X_i = \begin{cases} 1 \text{ si le } i^{\text{ème}} \text{ patient traité au médicament } m_1 \text{ a fait une thrombose} \\ 0 \text{ sinon} \end{cases}$$

et

$$Y_j = \begin{cases} 1 \text{ si le } j^{\text{ème}} \text{ patient traité au médicament } m_2 \text{ a fait une thrombose} \\ 0 \text{ sinon} \end{cases}$$

On suppose que les $(X_i)_{i=1..n_1}$ (resp. $(Y_j)_{j=1..n_2}$) sont des v.a indépendantes de Bernoulli de paramètre p_1 (resp. p_2). Les paramètres p_1 et p_2 représentent donc les probabilités respectives de faire au moins une thrombose dans l'année lorsqu'on est traité aux médicaments m_1 et m_2. Notre objectif est d'évaluer p_1 et p_2.

Rappelons maintenant quelques définitions et notations utiles dans la suite de l'exposé.

1.2 Définitions et notations

1.2.1 Variables aléatoires de Bernoulli, Gaussienne

Si X est une variable aléatoire intégrale, on note $\mathbb{E}(X)$ son espérance. Si elle est en plus de carré intégrable, on note $\mathrm{Var}(X)$ sa variance. Si sa loi est $\mathcal{L}$, on écrit $X \sim \mathcal{L}$.

Définition 1.2.1. Une variable aléatoire de Bernoulli de paramètre p avec $p \in [0;1]$ est une variable aléatoire réelle $X \sim \mathcal{B}(p)$, où $\mathcal{B}(p) = p\delta_1 + (1-p)\delta_0$.

Un calcul simple montre que $\mathbb{E}(X) = p$ et $\mathrm{Var}(X) = p(1-p)$.

Définition 1.2.2. Une variable aléatoire gaussienne réelle est une variable aléatoire X de densité $\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-m)^2}{2\sigma^2}\right)$ par rapport à la mesure de Lebesgue. On note $X \sim \mathcal{N}(m, \sigma^2)$ où $m \in \mathbb{R}$ est la moyenne de X et $\sigma^2 \geqslant 0$ sa variance.

Définition 1.2.3. On appelle modèle statistique la donnée d'un triplet $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ où Ω est un ensemble, $\mathcal{A}$ est une tribu et $(P_\theta)_{\theta\in\Theta}$ est une famille de probabilités sur $(\Omega, \mathcal{A})$. Par convention, toute quantité indicée par θ est définie sous la probabilité P_θ.

Dans l'exposé, p_1 et p_2 varient dans $[0;1]^2$, on considère donc $\Theta = [0;1]^2$. Les v.a $(X_i)_{i=1..n_1}$ i.i.d de loi $\mathcal{B}(p_1)$ et $(Y_j)_{j=1..n_2}$ i.i.d de loi $\mathcal{B}(p_2)$ sont à valeurs dans l'espace mesurable $(\{0,1\}, \mathcal{P}(\{0,1\}))$. On considère donc le modèle statistique suivant :

$$(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta}) = (\{0,1\}^{n_1+n_2}, \mathcal{P}(\{0,1\})^{n_1+n_2}, (P_{p_1,p_2} = \mathcal{B}(p_1)^{\otimes n_1} \otimes \mathcal{B}(p_2)^{\otimes n_2})_{(p_1,p_2)\in[0;1]^2})$$

1.3 Retour au problème

On note

$$\widehat{p_1} = \frac{\text{nombre de patients du groupe 1 qui ont subit une attaque}}{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$$

De même,

$$\widehat{p_2} = \frac{\text{nombre de patients du groupe 2 qui ont subit une attaque}}{n_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$$

$\widehat{p_1}$ et $\widehat{p_2}$ sont des estimateurs de p_1 et p_2.

Introduisons maintenant la notion d'estimateur :

Définition 1.3.1. (estimateur) Soit $g : \Theta \to \mathbb{R}^d$. On appelle estimateur de $g(\theta)$ au vu de l'observation $X = (X_1, ..., X_n)$ toute variable aléatoire $\sigma(X)$-mesurable $T : \Omega \to \mathbb{R}^d$. Si $\mathbb{E}_\theta(|T|) < \infty$, on appelle biais de l'estimateur la fonction

$$b_T : \Theta \to \mathbb{R}^d \theta \to \ \mathbb{E}_\theta(T) - g(\theta)$$

On dit que T est sans biais si $b_T = 0$.

$\widehat{p_1}$ et $\widehat{p_2}$ sont des estimateurs sans biais de p_1 et p_2 au vu de l'observation $Z = (X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2})$. En effet, $\widehat{p_1}$ et $\widehat{p_2}$ sont bien $\sigma(Z)$-mesurables, pour tout $(p_1, p_2) \in [0;1]^2$, $\widehat{p_1}$ et $\widehat{p_2}$ sont L^1 sous P_{p_1,p_2} et $\mathbb{E}_{p_1,p_2}(\widehat{p_1}) = p_1$, $\mathbb{E}_{p_1,p_2}(\widehat{p_2}) = p_2$.

Intuitivement, les estimateurs $\widehat{p_1}$ et $\widehat{p_2}$ sont des quantités qui approchent "bien" p_1 et p_2.

1.3.1 Test d'hypothèses

On se place maintenant dans le modèle statistique $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$. Soient Θ_0 et Θ_1 deux sous ensembles disjoints de Θ. En observant ω, on veut tester $\theta \in \Theta_0$ contre $\theta \in \Theta_1$. On appelle $H_0 : \theta \in \Theta_0$ l'hypothèse principale ou l'hypothèse nulle et $H_1 : \theta \in \Theta_1$ l'hypothèse alternative.

Définition 1.3.2. On appelle test toute variable aléatoire $\Delta : \Omega \to \{0,1\}$.

Un test s'utilise grâce à une règle de décision :

$$\text{On accepte } H_0 \Leftrightarrow \Delta(w) = 0$$
$$\text{On rejette } H_0 \Leftrightarrow \Delta(w) = 1$$

Remarque 1.3.3. Un test n'est pas symétrique. Un test de H_0 contre H_1 est différent d'un test de H_1 contre H_0. En effet, dans un test de H_0 contre H_1, on a un parti pris pour H_0 et on ne rejette H_0 que lorsqu'on ne peut pas faire autrement.

Décider qu'un médicament est meilleur que l'autre revient à évaluer la quantité $p_1 - p_2$. On ne privilégie a priori aucun des deux médicaments, on va donc faire un test de $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$. (Si on fait un test de $H_0 : p_1 \leqslant p_2$ contre $H_1 : p_1 > p_2$, on part avec un a priori positif pour le médicament m_1, de même un test de $H_0 : p_1 \geqslant p_2$ contre $H_1 : p_1 < p_2$ part avec un a priori positif pour le médicament m_2).
Pour discuter de la qualité d'un test on introduit plusieurs notions.

Définition 1.3.4. On dit que Δ est de niveau α si

$$\sup_{\theta \in \Theta_0} P_\theta(\Delta = 1) \leqslant \alpha$$

A priori, un test ou un estimateur dépend du nombre n d'observations. Ainsi on parle de niveau asymptotique α si

$$\sup_{\theta \in \Theta_0} (\lim_{n \to \infty} P_\theta(\Delta = 1)) \leqslant \alpha$$

Remarque 1.3.5. Un test est construit pour valider (ou infirmer) l'hypothèse H_0. En général, on se donne un niveau α, une hypothèse H_0 et on **construit** le test pour qu'il soit de niveau α. La notion de niveau est donc la caractéristique essentielle d'un test.

Définition 1.3.6. La fonction puissance d'un test Δ est $f : \Theta_1 \to [0, 1]$ définie par $f(\theta) = P_\theta(\Delta = 1)$.

Remarque 1.3.7. L'erreur de première espèce au point $\theta \in \Theta_0$ est la quantité $P_\theta(\Delta = 1)$ (la probabilité de rejeter H_0 sous $\theta \in \Theta_0$). L'erreur de deuxième espèce au point $\theta \in \Theta_1$ est la quantité $P_\theta(\Delta = 0)$ (la probabilité de rejeter H_1 sous $\theta \in \Theta_1$). Ces deux notions permettent de mieux cerner le test.

La puissance représente la probabilité de rejeter H_0 sous H_1. On veut donc imposer à la puissance d'être le plus proche possible de 1. Le niveau représente quant à lui une majoration de la probabilité de rejeter H_0 sous H_0, α doit donc être plutôt faible.

A niveau fixe, pour que le test soit "bon", on veut que la puissance soit la plus grande possible.

2 Etude du premier modèle

On se donne ici un niveau α (typiquement $\alpha = 5\%$) et on cherche à construire un test Δ de niveau α.
On considère un test Δ du type :

$$\Delta = \begin{cases} 0 \text{ si } |\widehat{p_1} - \widehat{p_2}| \leqslant d_{\alpha,n_1,n_2} \\ 1 \text{ sinon} \end{cases}$$

On cherche donc à déterminer le d_{α,n_1,n_2} qui assure que le test Δ est bien de niveau α.
D'après le Théorème Central Limite :

$$\sqrt{n_1} \frac{\widehat{p_1} - p_1}{\sqrt{p_1(1-p_1)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Comme les échantillons utilisés pour tester les médicaments m_1 et m_2 sont distincts, on peut supposer $\widehat{p_1}$ et $\widehat{p_2}$ indépendants. On a donc au final :

$$\begin{cases} \frac{\widehat{p_1}-p_1}{\sqrt{\frac{p_1(1-p_1)}{n_1}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \\ \frac{\widehat{p_2}-p_2}{\sqrt{\frac{p_2(1-p_2)}{n_2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \end{cases} \implies \frac{\widehat{p_1}-\widehat{p_2}-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

On connaît la loi asymptotique de $\widehat{p_1}-\widehat{p_2}$. On va donc essayer de construire un test de niveau asymptotique α (c.à.d $\lim\limits_{n_1,n_2\to\infty} P_{H_0}(\Delta=1) \leqslant \alpha$). Pour cela, il suffit de choisir d_{α,n_1,n_2} tel que, asymptotiquement, $P_{H_0}(|\widehat{p_1}-\widehat{p_2}| \geqslant d_{\alpha,n_1,n_2}) \leqslant \alpha$.
Comme on se place sous l'hypothèse H_0, on suppose $p_1 = p_2$. On a donc

$$\frac{\widehat{p_1}-\widehat{p_2}}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

On doit maintenant choisir d_{α,n_1,n_2} tel que $P\left(|\mathcal{N}(0,1)| \geqslant \frac{d_{\alpha,n_1,n_2}}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\right) \leqslant \alpha$. A priori d_{α,n_1,n_2} dépend de $p_1 = p_2$. Mais, sur $[0,1]$, $p_1(1-p_1) \leqslant \frac{1}{4}$ et $p_2(1-p_2) \leqslant \frac{1}{4}$. Donc $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \leqslant \sigma^2 = \frac{1}{4n_1}+\frac{1}{4n_2}$. On a réussi à majorer la variance de $\widehat{p_1}-\widehat{p_2}$ indépendamment de p_1 et p_2. Or, asymptotiquement

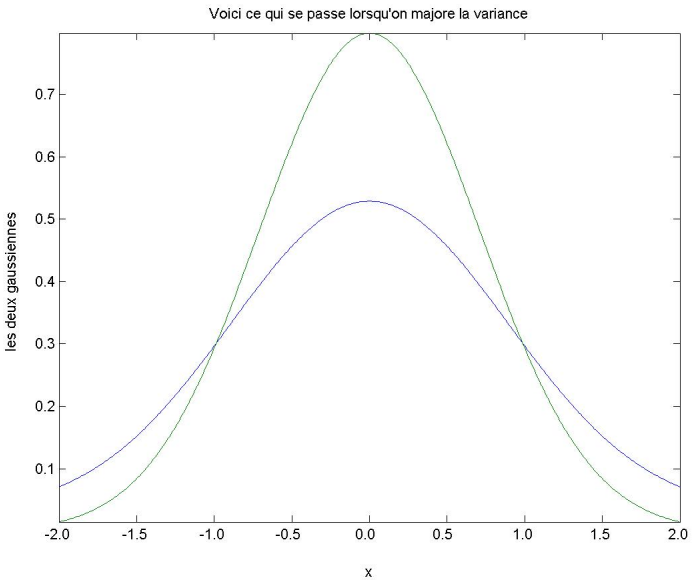
$$P(|\widehat{p_1}-\widehat{p_2}| \geqslant d_{\alpha,n_1,n_2}) = P\left(|\mathcal{N}(0,1)| \geqslant \frac{d_{\alpha,n_1,n_2}}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\right)$$

croît avec la variance $\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}$. Il suffit donc de la majorer pour la plus grande variance, σ^2.
Finalement, asymptotiquement

$$\forall p_1 \in [0;1], \forall p_2 \in [0;1], P(|\widehat{p_1}-\widehat{p_2}| \geqslant d_{\alpha,n_1,n_2}) \leqslant P\left(|\mathcal{N}(0,1)| \geqslant \frac{d_{\alpha,n_1,n_2}}{\sigma}\right)$$

Il suffit donc de choisir d_{α,n_1,n_2} tel que $P\left(|\mathcal{N}(0,1)| \geqslant \frac{d_{\alpha,n_1,n_2}}{\sigma}\right) = \alpha$. On note $u_{\frac{\alpha}{2}}$ le quantile d'ordre $\frac{\alpha}{2}$ de la gaussienne centrée réduite ($P(\mathcal{N}(0,1) \geqslant u_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$). Comme la gaussienne $\mathcal{N}(0,1)$ est symétrique, on a $d_{\alpha,n_1,n_2} = \sigma u_{\frac{\alpha}{2}} = u_{\frac{\alpha}{2}}\sqrt{\frac{1}{4n_1}+\frac{1}{4n_2}}$.

Remarque 2.0.8. On a ici construit un test de niveau asymptotique α. En fait, une pratique résultant de simulations donne mieux. Dès que $n_1p_1 \geqslant 5$ et $n_1(1-p_1) \geqslant 5$, les fonctions de répartitions de $\widehat{p_1}$ et $\mathcal{N}(p_1, \frac{p_1(1-p_1)}{n_1})$ ne diffèrent qu'au troisième chiffre après la virgule. D'après cette approximation, dès que n_1p_1 , $n_1(1-p_1) \geqslant 5$ et n_2p_2 , $n_2(1-p_2) \geqslant 5$, "$n_1 = n_2 = \infty$", i.e les probabilités $P(|\widehat{p_1}-\widehat{p_2}| \leqslant d_{\alpha,n_1,n_2})$ et $P\left(|\mathcal{N}(p_1-p_2, \frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2})| \geqslant d_{\alpha,n_1,n_2}\right)$ ne diffèrent qu'au troisième chiffre après la virgule.



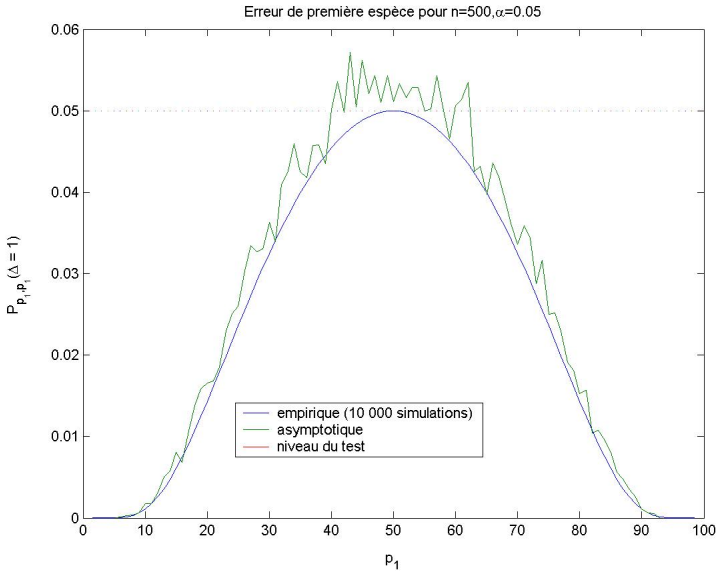
2.1 Niveau et puissance

Erreur de première espèce

Maintenant que le test Δ a été bien défini, on peut vérifier expérimentalement (à l'aide de simulations) qu'il est bien de niveau α. On se donne plusieurs valeurs de $p_1 = p_2$ et on calcule l'erreur empirique de première espèce en (p_1, p_1) (on espère trouver un résultat inférieur à α, le niveau asymptotique). Les résultats de ces simulations sont présentés dans la figure suivante.

On peut aussi calculer l'erreur asymptotique de première espèce en (p_1, p_1). Elle est donnée par $P_{p_1,p_1}(\Delta = 0) = P_{H_0}(|\mathcal{N}(0, p_1(1-p_1)(\frac{1}{n_1} + \frac{1}{n_2}))| \leqslant d_{\alpha,n_1,n_2}) = g(p_1)$. Le graphe de g est représenté sur la figure suivante.

Pour le d_{α,n_1,n_2} choisi, on a bien sûr $g(p_1) \leqslant \alpha$ mais, comme le montre la figure, on peut même avoir $g(p_1) \ll \alpha$ quand $p_1(1-p_1) \ll \frac{1}{4}$. La majoration est donc assez grossière.



Remarque 2.1.1. Comme prévu, l'erreur empirique de première espèce ne diffère de l'erreur asymptotique de deuxième espèce qu'au troisième chiffre après la virgule.

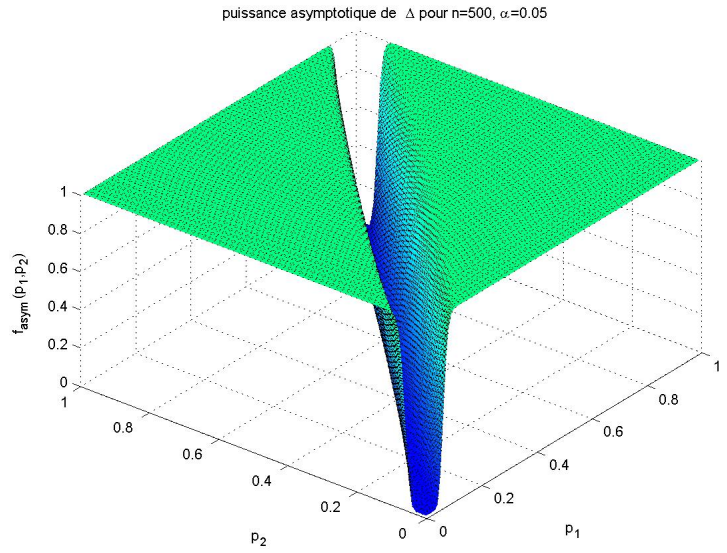
Puissance du test

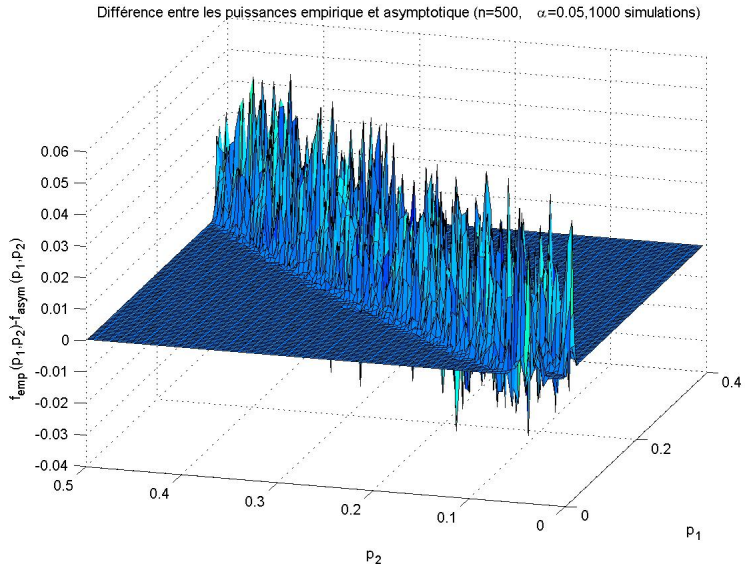
On a défini un test Δ de niveau α, efficace pour valider H_0 sous H_0. On s'intéresse maintenant à la puissance de Δ (qui représente la capacité du test Δ à rejeter H_0 sous H_1).
Comme dans la partie précédente, on a

$$\frac{\widehat{p_1} - \widehat{p_2} - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Donc $f_{asym}(p_1, p_2) = P_{p_1,p_2}(\Delta_1 = 1) = P(|\mathcal{N}(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2})| \leqslant d_{\alpha,n_1,n_2})$

On a trouvé une formule explicite pour la puissance asymptotique, mais on peut, de même que pour le niveau, calculer une puissance empirique. On espère que les deux fonctions puissance ne sont pas trop éloignées (ce qui justifie a posteriori l'approximation gaussienne).





La première figure représente le graphe de la puissance asymptotique du test Δ. La seconde représente la différence entre les puissances asymptotique et empirique (10000 simulations). La différence est inférieure à 0.05 en valeur absolue.

Remarque 2.1.2. Les simulations montrent que les différences entre puissances asymptotique et empirique ne sont pas significatives. On s'intéresse donc désormais uniquement aux formules asymptotiques. Les calculs faits dans la suite de cette partie seront donc tous faits à partir des formules asymptotiques.

Remarque 2.1.3. Il est important de noter que la "diagonale" de la puissance représente l'erreur de première espèce ($f(p_1, p_1)$ représente l'erreur de première espèce en p_1). Mieux le test est

calibré, meilleure est la puissance. En effet, si on choisit d_α de sorte que $P_{H_0}(|\widehat{p_1} - \widehat{p_2}| \geqslant d_\alpha) \leqslant \alpha$, on peut définir un test de niveau α en posant

$$\Delta = \begin{cases} 0 \text{ si } |\widehat{p_1} - \widehat{p_2}| \leqslant d_\alpha \\ 1 \text{ sinon} \end{cases}$$

Si on choisit d'_α tel que $P_{H_0}(|\widehat{p_1} - \widehat{p_2}| \geqslant d'_\alpha) = \alpha$ (i.e on prend le plus grand d_α possible), on peut définir de manière analogue un test Δ' de niveau α mais surtout

$$f'(p_1, p_2) = P_{p_1,p_2}(|\widehat{p_1} - \widehat{p_2}| \geqslant d'_\alpha) \geqslant P_{p_1,p_2}(|\widehat{p_1} - \widehat{p_2}| \geqslant d_\alpha) = f(p_1, p_2)$$

d'_α donne une meilleure puissance que d_α. Il est donc important de définir le niveau au plus juste.

On voit que, hormis sur la diagonale, la puissance est très proche de 1. On se donne une marge de sécurité β et on peut définir deux berges à $1 - \beta$ par $B = \{(p_1, p_2) : f(p_1, p_2) \geqslant 1 - \beta\}$. On pose ensuite $e_\beta = \sup_{p_1} \inf_{p_2} \{|p_1 - p_2| : (p_1, p_2) \in B\}$. e_β est l'écart minimum qu'on peut détecter avec une probabilité supérieure à $1 - \beta$ (i.e si $|p_1 - p_2| \geqslant e_\beta$, on rejette l'hypothèse $H_0 : p_1 = p_2$ avec une probabilité $f(p_1, p_2) \geqslant 1 - \beta$).

2.2 Inversion de la figure

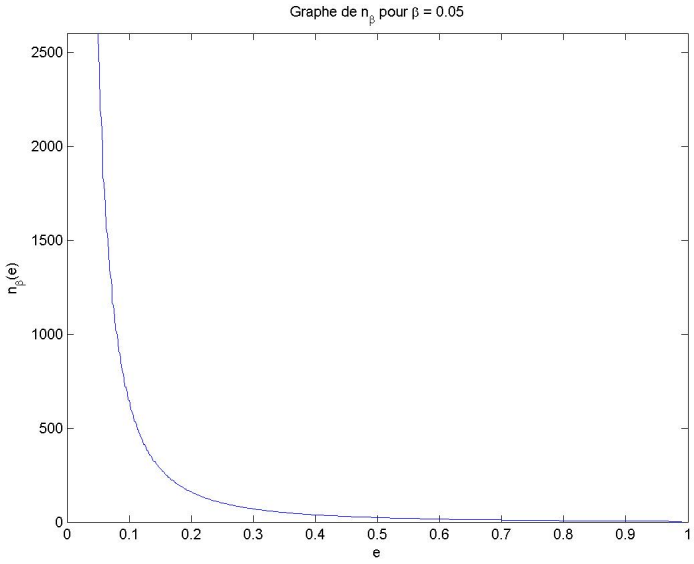
On suppose maintenant $n_1 = n_2 = n$. A n et β donnés, on peut donc détecter un écart $e_\beta(n)$ avec une marge de sécurité β. On se propose "d'inverser" cette courbe : on se donne β et e et on cherche le plus petit n tel que $e = e_\beta(n)$ (c'est-à-dire le nombre minimum de patients dont on a besoin pour distinguer p_1 et p_2 à e près avec une marge de sécurité β). On note $\sigma^2(p_1, p_2) = p_1(1 - p_1) + p_2(1 - p_2)$. On a

$$f_{n,app}(p_1, p_2) = 1 - P(|\mathcal{N}(p_1 - p_2, \frac{\sigma^2(p_1, p_2)}{n})| \geqslant d_{\alpha,n}) = 1 - P(|\mathcal{N}(0, 1) - \frac{\sqrt{n}(p_2 - p_1)}{\sigma(p_1, p_2)}| \leqslant \frac{\sqrt{n} d_{\alpha,n}}{\sigma(p_1, p_2)})$$

Comme $\frac{\sqrt{n} d_{\alpha,n}}{\sigma(p_1, p_2)} = \frac{u_{\frac{\alpha}{2}}}{\sqrt{2}\sigma(p_1, p_2)} = C$, on a en fait

$$f_{n,app}(p_1, p_2) = 1 - P(|\mathcal{N}(0, 1) - \frac{\sqrt{n}(p_2 - p_1)}{\sigma(p_1, p_2)}| \leqslant C)$$

On a donc $f_{n,app}(p_1, p_2) \longrightarrow_{n \to \infty} 1$ dès que $p_1 \neq p_2$.
On pose donc $n_\beta(e) = \sup_{|p_1 - p_2| \geqslant e} \{\inf\{n : f_{n,app}(p_1, p_2) \geqslant 1 - \beta\}\}$.



3 Etude du deuxième modèle

Le modèle présenté ci-dessus possède un certain nombre de défauts. En pratique, on ne dispose pas de tous les patients dès le début de l'expérience, certains la rejoignent en cours de route. Ce fait n'est pas pris en compte dans notre modèle. De plus, pour valider ou rejeter H_0, il faut attendre la fin de l'expérience. Enfin, on a majoré la variance très grossièrement (par $\frac{1}{4}$, ce qui correspond à $p_1 = p_2 = 0.5$ alors que les données médicales donnent $p_1 \simeq p_2 \simeq 0.10$), on peut donc espérer améliorer la puissance en recalibrant le test. On aimerait un modèle plus souple qui réponde à ces trois exigences : de nouveaux patients peuvent rejoindre l'expérience en cours de route, on peut arrêter l'expérience avant son terme si on a détecté un écart significatif entre p_1 et p_2, on aimerait majorer la variance au plus juste.

3.1 Approximation Poissonnienne

Une façon naturelle de définir la nouvelle expérience est la suivante : on augmente la fréquence des MMT (au lieu de remplir un MMT par an, le patient en remplit un toutes les deux semaines) et on sort de l'expérience dès qu'on détecté un écart significatif entre p_1 et p_2.

Dans cette nouvelle expérience, on ne s'intéresse plus au paramètre n (le nombre de patients impliqués dans l'expérience) mais au paramètre N, le nombre total de MMT remplis par l'ensemble des patients depuis qu'ils ont rejoint l'expérience. En fait, on s'intéresse toujours au nombre total d'observations mais il ne coïncide plus avec le nombre de patients.

On cherche à construire un bon test (de niveau α et avec la plus grande puissance possible). On cherche donc à évaluer la probabilité p' qu'un patient fasse au moins une thrombose en deux semaines en fonction de la probabilité p d'en faire au moins une en un an. L'estimation de p' permettra de majorer la variance $p'(1-p')$ au mieux, et de calibrer le test au plus juste .

Définition 3.1.1. Un processus de Poisson de paramètre λ ($\lambda > 0$) est un processus de comptage qui vérifie :

(i) $N(0) = 0$
(ii) Le processus est à incréments indépendants.
(iii) Le nombre d'événements qui intervient dans n'importe quel intervalle de longueur t suit une loi de Poisson de paramètre λt. Pour tout $s, t \geqslant 0$, pour tout $n \in \mathbb{N}$,

$$P(N(s+t) - N(s) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

Définition 3.1.2. Un processus de comptage $(N(t))_{t \geqslant 0}$ est à incréments stationnaires si le nombre d'événements qui interviennent dans un intervalle de temps ne dépend que de la taille de cet intervalle de temps. Autrement dit, pour tous $s, t \geqslant 0$, la loi de $N(s+t) - N(s)$ ne dépend que de t (et pas de s).

Définition 3.1.3. Un processus de comptage $(N(t))_{t \geqslant 0}$ est à incréments indépendants si le nombre d'événements qui interviennent dans des intervalles de temps disjoints sont indépendants : pour tous $p \leqslant q \leqslant r \leqslant s$, $N(s) - N(r)$ et $N(q) - N(p)$ sont indépendants.

On introduit une autre définition d'un processus de Poisson, équivalente à la première mais plus pratique dans le cas qui nous intéresse.

Définition 3.1.4. Un processus de Poisson de paramètre λ ($\lambda > 0$) est un processus de comptage qui vérifie :
(i) $N(0) = 0$
(ii) Le processus est à incréments indépendants et stationnaires.
(iii) $P(N(h) = 1) = \lambda h + o(h)$
(iv) $P(N(h) \geqslant 2) = o(h)$

Remarque 3.1.5. Si $(N(t))_{t \in \mathbb{R}}$ est un processus de Poisson de paramètre λ, on a

$$\mathbb{E}(N(t)) = \lambda t$$

On peut voir λ comme la fréquence d'apparition d'un événement.

On aimerait modéliser le nombre de thromboses par un processus de comptage simple : le processus de Poisson. Justifions l'approximation Poissonnienne. On commence l'observation d'un patient à la date 0 et on note $N(t)$ le nombre de thromboses faites par le patient entre les dates 0 et t. On a évidemment $N(0) = 0$. Les données médicales montrent que le nombre de thromboses qui surviennent durant un intervalle de temps donné ne dépend pas du nombre de thromboses survenues avant cet intervalle mais uniquement de la longueur de cet intervalle de temps : autrement dit, les incréments sont stationnaires et indépendants. Les thromboses n'arrivent pas par "paquets", l'hypothèse $P(N(t) \geqslant 2) = o(h)$ est donc raisonnable. Enfin, si on introduit la fréquence d'apparition f des thromboses, durant une durée t, il se produit en moyenne ft thromboses. Il apparaît donc raisonnable de supposer $P(N(t) = 1) = ft + o(t)$ pour t petit.

Remarque 3.1.6. Bien entendu l'approximation Poissonnienne n'est qu'une approximation. En pratique, le nombre de thromboses est influencé par un certain nombre de facteurs, dont certains saisonniers, qui font que les incréments ne sont ni stationnaires ni indépendants. D'autres processus, plus complexes, peuvent prendre en compte la variabilité de la fréquence : le processus de Poisson inhomogène par exemple. Cependant, pour garder un modèle simple, on ne tiendra pas compte de ces facteurs.

Dans notre problème, on passe d'un test par an à un test toutes les deux semaines. $p = P(N(365) \geqslant 1) = 1 - e^{-365\lambda}$ et $p' = P(N(14) \geqslant 1) = 1 - e^{-14\lambda}$. Donc $p' = 1 - e^{\frac{14}{365} \ln(1-p)}$. Le calcul avec $p = 0.10$ donne $p' = 0.004$.

Remarque 3.1.7. Une autre manière, plus simple, de calculer p' est de faire le raisonnement suivant : $1-p'$ est la probabilité de ne pas faire de thrombose en deux semaines, $1-p$ est la probabilité de ne pas faire de thrombose en un an. Comme les événements "ne pas faire de thrombose en deux semaines" sont indépendants pour des séquences disjointes de deux semaines (incréments indépendants) et qu'un an comporte 26 séquences disjointes de deux semaines, $(1-p) = (1-p')^{26}$ donc $p' = 1-(1-p)^{\frac{1}{26}}$. Le calcul avec $p = 0.10$ donne $p' = 0.004$ (il est normal de retrouver la même valeur qu'avec l'approximation Poissonnienne).

3.2 Première approche

On tente de se ramener au modèle précédent. Voici la démarche adoptée dans un premier temps.

Pour simplifier les calculs, on suppose $n_1 = n_2 = n$ (on a le même nombre d'observations pour chaque médicament) et on considère les v.a $(X_i)_{i=1..n}$ et $(Y_j)_{j=1..n}$ définies par :

$$X_i = \begin{cases} 1 \text{ si la } i^{\text{ème}} \text{ observation pour le médicament } m_1 \text{ révèle une thrombose} \\ 0 \text{ sinon} \end{cases}$$

et

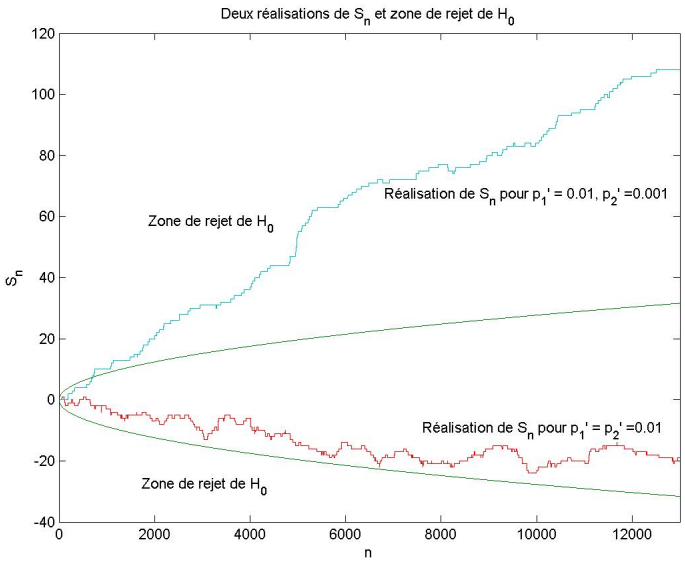
$$Y_j = \begin{cases} 1 \text{ si la } j^{\text{ème}} \text{ observation pour le médicament } m_2 \text{ révèle une thrombose} \\ 0 \text{ sinon} \end{cases}$$

Comme précédemment, on suppose que les $(X_i)_{i=1..n}$ (resp. les $(Y_j)_{j=1..n}$) sont des v.a indépendantes de Bernoulli de paramètre p'_1 (resp. p'_2).
On considère ensuite la suite de v.a $S_n = \sum_{i=1}^{n} X_i - Y_i$. On a

$$\frac{S_n - (p'_1 - p'_2)}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, p'_1(1-p'_1) + p'_2(1-p'_2))$$

Comme dans la partie précédente, dès que np'_1 , $np'_2 \geqslant 5$ et $n(1-p'_1)$, $n(1-p'_2) \geqslant 5$, on peut faire l'approximation Gaussienne. On peut donc calibrer t_α pour que $P_{H_0}\left(\left|\frac{S_n}{\sqrt{n}}\right| \geqslant t_\alpha\right) \leqslant \alpha$ pour tous les n qui permettent de faire l'approximation Gaussienne. On peut donc rejeter l'hypothèse nulle $H_0 : p'_1 = p'_2$ dès qu'on trouve un n assez grand tel que $\left|\frac{S_n}{\sqrt{n}}\right| \geqslant t_\alpha$. De manière formelle, on pose $n_{min} = \inf\{n : np'_1, np'_2 \geqslant 5 \text{ et } n(1-p'_1), n(1-p'_2) \geqslant 5\}$ et on considère le test :

$$\Delta = \begin{cases} 0 \text{ si } \sup\limits_{n=n_{min}..n_{max}} \left|\frac{S_n}{\sqrt{n}}\right| \leqslant t_\alpha \\ 1 \text{ sinon} \end{cases}$$



La figure ci-dessus montre ce qu'on fait en pratique : on calcule S_n pour tout n et on vérifie que $\left|\frac{S_n}{\sqrt{n}}\right| \leqslant t_\alpha$ (i.e S_n reste dans la parabole $y^2 = t_\alpha^2 x$). On a choisi $n_{max} = 13000$ qui correspond à 500 patients suivis toutes les deux semaines pendant un an et on a représenté deux réalisations de S_n : une pour $p_1' = p_2' = 0.01$, qui ne sort pas de la parabole et une pour $p_1' = 0.01$, $p_2' = 0.001$ qui sort de la parabole pour $n = 707$ (soit bien avant la fin de l'expérience).

Remarque 3.2.1. Le sup indique qu'on calcule S_n pour tous les n (d'où le nom de test séquentiel). Dès qu'on trouve un S_n qui vérifie $\left|\dfrac{S_n}{\sqrt{n}}\right| \geqslant t_\alpha$, le sup vérifie aussi la condition donc on peut sortir de l'expérience sans plus attendre.
On majore aussi la variance plus finement. Les données médicales fournissent $p_1' \simeq p_2' \simeq 0.0043$. On majore p_1' , $p_2' \leqslant 0.01$ et la variance $p_1'(1-p_1')$, $p_2'(1-p_2') \leqslant \sigma^2 = 0.01$ (au lieu de $\frac{1}{4}$ dans le premier modèle). L'approximation gaussienne

$$P_{H_0}\left(\left|\frac{S_n}{\sqrt{n}}\right| \geqslant t_\alpha\right) = \sup_{p_1'} P(|\mathcal{N}(0, 2p_1'(1-p_1')| \geqslant t_\alpha)$$

donne $t_\alpha = u_{\frac{\alpha}{2}}\sqrt{2}\sigma = 0.2772$.
Enfin, dans ce modèle, on s'intéresse uniquement au nombre total d'observations, les patients n'ont pas besoin d'être là dès le début, ils peuvent rejoindre et quitter l'expérience en cours de route. Le modèle répond donc aux exigences qu'on attend de lui.

En fait, à cause de la loi du log itéré, ce modèle ne fonctionne pas. De manière asymptotique, on rejette systématiquement l'hypothèse H_0 sous H_0.

Théorème 3.2.2. *Si $(B_t)_{t\geqslant 0}$ est un mouvement brownien alors*

$$p.s \quad \limsup_{t\to\infty} \frac{B_t}{\sqrt{2t\ln\ln t}} = 1$$

Preuve 1 :
Lemme : On considère le temps d'arrêt $T_a = \inf\{t : B_t = a\}$

$$P(\max_{0\leqslant s\leqslant 1} B_s \geqslant a) = P(T_a \leqslant 1) = 2P(B_1 \geqslant a) \quad (1)$$

$$\int_x^\infty e^{-y^2/2}dy \leqslant \frac{e^{-x^2/2}}{x} \quad (2)$$

$$\int_x^\infty e^{-y^2/2}dy \sim \frac{e^{-x^2/2}}{x} \quad (3)$$

(1) est une conséquence du principe de réflexion pour un mouvement brownien. On utilise simplement que $\max_{0\leqslant s\leqslant t} B_s$ a même loi que $\mid B_t \mid$.

(2) est immédiat en majorant $e^{-y^2/2}$ par $\frac{y}{x}e^{-y^2/2}$ dans l'intégrale.

1) Montrons pour commencer

$$p.s \quad \limsup_{t\to\infty} \frac{B_t}{\sqrt{2t\ln\ln t}} \leqslant 1$$

Soit $\alpha > 1$ et $t_n = \alpha^n$

$$\begin{aligned} P(\max_{t_n\leqslant s\leqslant t_{n+1}} B_s > (t_n f(t_n))^{1/2}) &\leqslant P(\max_{0\leqslant s\leqslant t_{n+1}} \frac{B_s}{t_{n+1}^{1/2}} > (\frac{f(t_n)}{\alpha})^{1/2}) \\ &\leqslant 2(2\pi)^{-1/2}(\tfrac{f(t_n)}{\alpha})^{1/2}\exp(\tfrac{-f(t_n)}{2\alpha}) \end{aligned}$$

d'après (1) et (2)
Si $f(t) = 2\alpha^2 \ln\ln t$ alors $\ln\ln t = \ln n \ln\alpha = \ln n + \ln\ln\alpha$ donc $\exp(-f(t_n)/2\alpha) \leqslant C_\alpha n^{-\alpha}$ où C_α ne dépend que de α.
D'où

$$\sum_{n=0}^{\infty} P\big(\max_{t_n\leqslant s\leqslant t_{n+1}} B_s > (t_n f(t_n))^{1/2}\big) \leqslant \infty$$

Par le lemme de Borel-Cantelli on a alors $P\big(\limsup_{n\to\infty}\big\{\max_{t_n\leqslant s\leqslant t_{n+1}} B_s > (t_n f(t_n))^{1/2}\big\}\big) = 0$
donc puisque $t \to (tf(t))^{1/2}$ est croissante on en déduit

$$P\big(\limsup_{n\to\infty}\big\{\max_{t_n\leqslant s\leqslant t_{n+1}} B_s > (sf(s))^{1/2}\big\}\big) = 0$$

puis

$$P\big(\limsup_{n\to\infty}\big\{\max_{t_n\leqslant s\leqslant t_{n+1}} \frac{B_s}{\sqrt{2s\ln\ln s}} > 1\big\}\big) = 0$$

on en déduit facilement que $p.s \quad \limsup_{t\to\infty} \frac{B_t}{\sqrt{2t\ln\ln t}} \leqslant 1$
2) Montrons maintenant l'autre sens de l'inégalité. On va cette fois utiliser la réciproque du lemme de Borel-Cantelli.
On prend à nouveau $\alpha \geqslant 1$ et $t_n = \alpha^n$. On s'intéresse à la quantité

$$P\Big(B_{t_{n+1}} - B_{t_n} \geqslant (t_{n+1}f(t_{n+1}))^{1/2}\Big) = P\Big(B_1 \geqslant (\beta f(t_{n+1}))^{1/2}\Big)$$

où $\beta = t_{n+1}/(t_{n+1} - t_n) = \alpha/(\alpha - 1) \geqslant 1$. D'après (3) si n est assez grand on peut majorer la quantité précédente par

$$\frac{1}{2\sqrt{2\pi}}(\beta f(t_{n+1}))^{-1/2} e^{-\beta f(t_{n+1})/2}$$

Si $f(t) = (2/\beta^2)\ln\ln t$ alors puisque $\ln\ln t_n = \ln n + \ln\ln\alpha$ on a

$$e^{-\beta f(t_{n+1})/2} \geqslant C_\alpha n^{-\beta/2}$$

où C_α ne dépend que de α. Finalement si $\alpha > 2$

$$\sum_{n=0}^{\infty} P\Big(B_{t_{n+1}} - B_{t_n} \geqslant (t_{n+1} f(t_{n+1}))^{1/2}\Big) = \infty$$

Les événements ci-dessus étant indépendants on peut utiliser la réciproque du lemme de Borel-Cantelli et déduire facilement que :

$$p.s \qquad \limsup_{n\to\infty} \frac{B_{t_{n+1}} - B_{t_n}}{((2/\beta^2) t_{n+1} \ln\ln t_{n+1})^{1/2}} \geqslant 1$$

d'où

$$p.s \qquad \limsup_{n\to\infty} \frac{B_{t_{n+1}} - B_{t_n}}{(2t_{n+1} \ln\ln t_{n+1})^{1/2}} \geqslant \frac{1}{\beta} = \frac{\alpha - 1}{\alpha}$$

$$p.s \qquad \limsup_{n\to\infty} \frac{B_{t_{n+1}}}{(2t_{n+1} \ln\ln t_{n+1})^{1/2}} \geqslant \frac{\alpha - 1}{\alpha} - \limsup_{n\to\infty} \frac{B_{t_n}}{(2t_{n+1} \ln\ln t_{n+1})^{1/2}} \geqslant \frac{\alpha - 1}{\alpha} - \frac{1}{\alpha^{1/2}}$$

Finalement $p.s \qquad \limsup_{t\to\infty} \frac{B_t}{(2t \ln\ln t)^{1/2}} \geqslant \frac{\alpha - 1}{\alpha} - \frac{1}{\alpha^{1/2}}$

En faisant $\alpha \to \infty$ on obtient bien le résultat voulu.

On déduit du théorème la proposition suivante :

Proposition 3.2.3. *Soit X_n une suite de v.a i.i.d qui vérifient $\mathbb{E}(X_1) = 0$ et $\mathrm{Var}(X_1) = 1$. On note $S_n = \sum_{i=1}^n X_i$, alors p.s :*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log(\log(n))}} = 1$$

Nous pouvons donc déduire de la proposition précédente que

$$p.s. \quad \limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} = \infty$$

et donc que

$$p.s. \quad \limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} \geqslant t_\alpha$$

C'est précisément le fait d'effectuer de façon répétée au cours du temps notre test de H_0 contre H_1 plutôt que de le faire une unique fois à la fin de l'expérience qui pose problème et rend le test inopérant. Il faut ajuster le modèle pour tenir compte de ce fait.

3.3 Ajustement du modèle

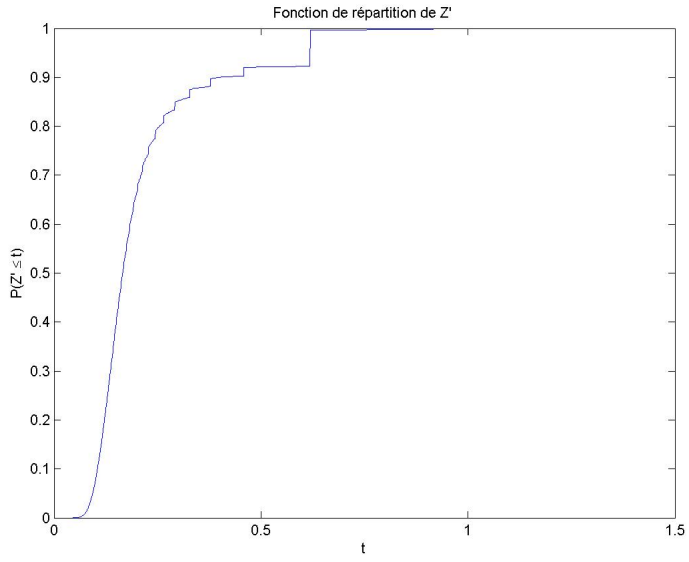
On considère maintenant la v.a $Z_{n_{max}} = \sup_{i=50..n_{max}} |Z_n| = \sup_{n=50..n_{\max}} \left| \frac{\sum_{i=1..n} X_i - Y_i}{\sqrt{2n \ln(\ln(n))}} \right|$.

On considère un test Δ du type :

$$\Delta = \begin{cases} 0 \text{ si } Z_{n_{max}} \leqslant v_\alpha \\ 1 \text{ sinon} \end{cases}$$

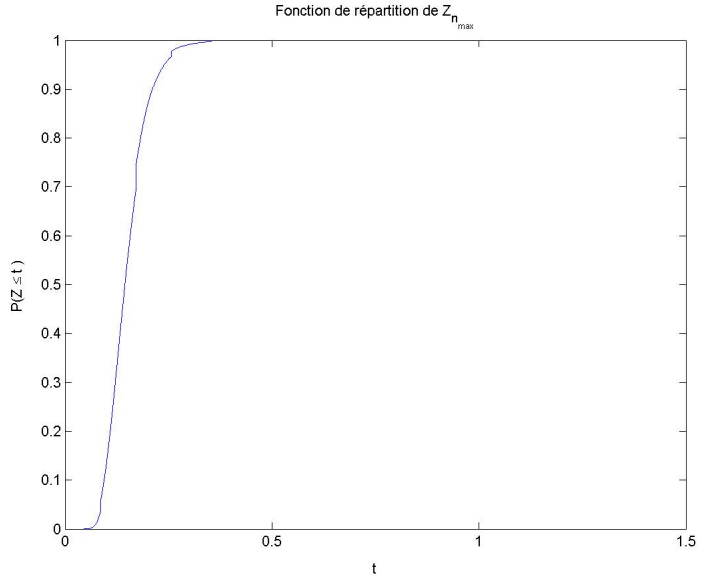
Et on cherche à calibrer v_α pour que le test soit de niveau α. A priori, v_α dépend encore une fois de $p_1 = p_2$. Il semble intuitif que v_α décroisse avec la variance mais on ne sait pas le prouver. On va donc calibrer v_α pour la variance maximale, c'est-à-dire pour $p_1' = p_2' = 0.01$.

Remarque 3.3.1. Pour des raisons pratiques on considère le sup sur $i \geqslant 50$. En effet, si on considère $Z' = \sup_{n=1..n_{\max}} \left| \frac{\sum_{i=1..n} X_i - Y_i}{\sqrt{2n \ln(\ln(n))}} \right|$, la fonction de répartition de Z' pose problème.



La présence d'atomes (dûe aux valeurs de Z_n pour les petites valeurs de n) ne permet pas de calibrer un test de niveau, par exemple, 0.05. La présence d'un atome au niveau du quantile c d'ordre 0.05 de Z' permet de définir un test de niveau 0.075 (on accepte H_0 si $Z' \leqslant c$) et un autre de niveau 0.002 (on accepte H_0 si $Z' < c$) mais rien entre les deux : on ne peut pas calibrer au plus juste un test de niveau 0.05.

On adopte la v.a $Z_{n_{max}}$ qui a une fonction de répartition plus "lisse".



Remarque 3.3.2. Le fait de considérer $Z_{n_{max}}$ au lieu de Z' revient à attendre d'avoir 50 observations pour commencer les tests. Ce n'est pas une hypothèse très contraignante. En effet, il est peu raisonnable de commencer un test statistique avec un nombre trop faible d'observations.

Maintenant que la fonction de répartition de $Z_{n_{max}}$ est connue, on peut calibrer le test au niveau qu'on veut. Dans la suite de l'exposé, on prend $\alpha = 0.05$, la fonction de répartition donne $v_\alpha = 0.2402$ (à comparer avec $\frac{t_\alpha}{\sqrt{2}} = 0.1960$ pour estimer la déviation dûe au $\log(\log(n))$)

Remarque 3.3.3. Il est tout de même possible de calculer un v_α théorique en faisant une majoration a priori assez grossière :

$$\begin{aligned}
P(\sup_{N_{min}\leqslant n\leqslant N_{max}} \frac{S_n}{\sqrt{2n\ln\ln n}} \geqslant v_\alpha) &\leqslant \sum_{n=N_{min}}^{N_{max}} P(\frac{S_n}{\sqrt{2n\ln\ln n}} \geqslant v_\alpha) \\
&\leqslant \sum_{n=N_{min}}^{N_{max}} P(\frac{S_n}{\sqrt{n}} \geqslant v_\alpha\sqrt{2\ln\ln n}) \\
&\leqslant \sum_{n=N_{min}}^{N_{max}} P(\mathcal{N}(0,\sigma^2) \geqslant v_\alpha\sqrt{2\ln\ln N_{min}}) \\
&\leqslant (N_{max} - N_{min})P(\mathcal{N}(0,\sigma^2) \geqslant v_\alpha\sqrt{2\ln\ln N_{min}})
\end{aligned}$$

En prenant $N_{min} = 50$ et $N_{max} = 13000$ et en majorant σ^2 par 0.01 on obtient $v_\alpha = 0.2703$ alors que nos simulations nous donnent $v_\alpha = 0.2403$. Le gain obtenu en faisant les simulations est finalement assez faible par rapport à ce qu'on peut trouver à l'aide de majorations brutales. Cependant, comme on veut la plus grande puissance possible, on doit calibrer le test au mieux et donc à prendre le plus petit v_α qui marche. On a donc intérêt à prendre le v_α donnée par les simulations.

3.4 Niveau et puissance

Maintenant que le test est défini, on s'intéresse aux erreurs de première et deuxième espèce. Contrairement au test précédent, les erreurs asymptotiques ne nous intéressent pas. En effet, comme les $X_i - Y_i - (p'_1 - p'_2)$ sont des v.a i.i.d de moyenne nulle et de variance $p'_1(1-p'_1) + p'_2(1-p'_2)$, d'après la loi du log itéré :

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} X_i - Y_i - n(p'_1 - p'_2)}{\sqrt{p'_1(1-p'_1) + p'_2(1-p'_2)}} \frac{1}{\sqrt{2n\log(\log(n))}} = 1$$

Notre majoration brutale donne :

$$v_\alpha \leqslant \frac{u_{\frac{\alpha}{n_{max}-n_{min}}}}{\sqrt{2\log(\log(n_{min}))}}$$

Or

$$\frac{\alpha}{n_{max}-n_{min}} = P(\mathcal{N}(0,1) \geqslant u_{\frac{\alpha}{n_{max}-n_{min}}}) \sim \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\left(-u_{\frac{\alpha}{n_{max}-n_{min}}}\right)^2}{2}\right)$$

Donc

$$u_{\frac{\alpha}{n_{max}-n_{min}}} \sim \sqrt{2\log(n_{max})}$$

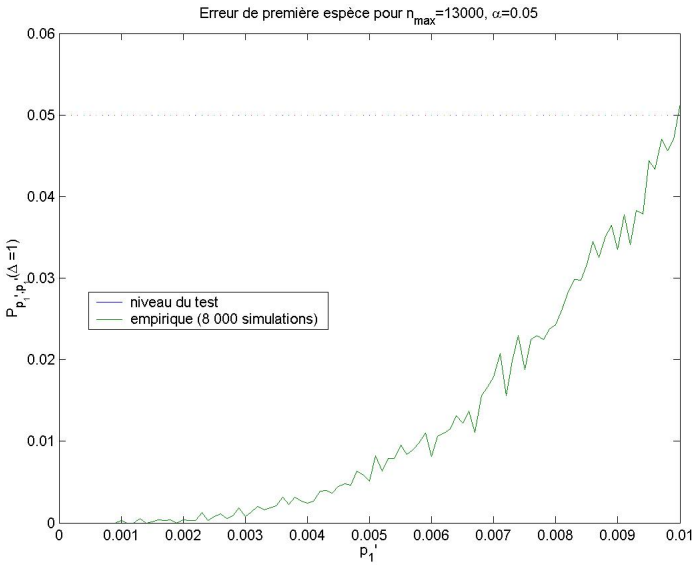
Si $p'_1 \neq p'_2$, p.s

$$Z_{n_{max}} \sim \left|\frac{n_{max}(p'_1 - p'_2)}{\sqrt{p'_1(1-p'_1) + p'_2(1-p'_2)}}\right| \frac{1}{\sqrt{2n_{max}\log(\log(n_{max}))}}$$

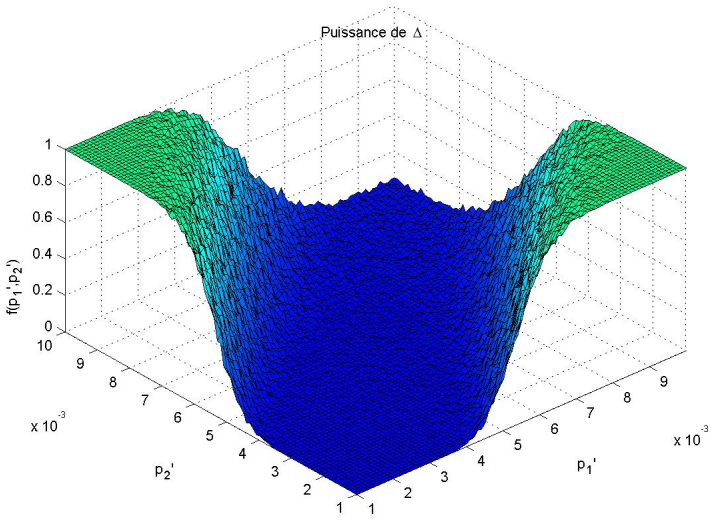
Donc p.s $\lim_{n_{max}\to\infty} \frac{Z_{n_{max}}}{v_\alpha} = \infty$. Finalement

$$\lim_{n_{max}\to\infty} P_{p'_1,p'_2}(\Delta = 1) = \lim_{n_{max}\to\infty} P_{p'_1,p'_2}(Z_{n_{max}} \geqslant v_\alpha) = 1$$

La puissance asymptotique vaut 1. On s'intéresse donc uniquement à la puissance empirique.



Remarque 3.4.1. Comme le montre la figure précédente, le test est bien de niveau $\alpha = 0.05$. Cependant, on avait majoré $p'_1 \leqslant 0.01$ et $p'_1(1-p'_1) \leqslant 0.01$. On voit que la majoration est peut-être encore un peu trop grossière puisque pour p'_1 de l'ordre de 0.005 (un peu plus que la valeur médicale), l'erreur de première espèce est de l'ordre de 0.005 soit 10 fois moins que le niveau du test.



Remarque 3.4.2. La puissance a l'air a priori moins bonne (les berges sont plus éloignées) dans le test séquentiel que dans le test précédent. Cependant, on considère ici des paramètres p'_1 et p'_2 faibles, les berges sont donc éventuellement plus éloignées uniquement à cause du changement d'échelle. Avec cette puissance, on peut détecter un écart de 0.006 avec une marge de sécurité $\beta = 0.05$. On cherche à comparer cette valeur à l'écart détectable avec la même marge que dans le premier test. Pour cela, il faut ramener cet écart entre des probabilités sur deux semaines à un écart entre des probabilités annuelles.

$$|p'_1 - p'_2| \geqslant 0.006 \Rightarrow f(p'_1, p'_2) \geqslant 1 - \beta$$

Du fait de l'approximation Poissonnienne, $|p'_1 - p'_2| = |(1 - e^{-14\lambda_1}) - (1 - e^{-14\lambda_2})|$. Compte tenu de l'ordre de grandeur de p'_1 et p'_2, on peut linéariser. Donc

$$|p'_1 - p'_2| \geqslant 0.006 \Leftrightarrow |\lambda_1 - \lambda_2| \geqslant \frac{0.006}{14} \Leftrightarrow |(1 - e^{-365\lambda_1}) - (1 - e^{-365\lambda_2})| \geqslant 365|\lambda_1 - \lambda_2| = 0.1590$$

On peut détecter un écart annuel de 0.1590 avec 13000 observations (soit 500 patients observés toutes les deux semaines pendant un an). Dans le premier test, avec 500 patients, on pouvait détecter un écart de 0.0723. La puissance du nouveau test est donc effectivement beaucoup moins bonne.

Conclusion

Nous avons présenté dans cet exposé un premier test statistique très élémentaire et les outils mathématiques qui permettent de rendre un tel test rigoureux puis nous avons tenté de mettre

en place une variante de ce premier test qui répond aux exigences du Dr Abastado. Plusieurs enseignements sont à tirer de ce travail :

D'abord les statistiques sont une discipline qui nécessite parfois de faire des simulations informatiques afin de calibrer au plus juste les tests mis en place. Cette pratique, surprenante a priori dans un exposé de mathématiques, est rendue inévitable par l'absence de résultats théoriques sur les vitesses de convergence de certains estimateurs.

Ensuite on constate une fois de plus à quel point il est délicat d'adapter des résultats théoriques à un problème pratique. En effet si la première méthode que nous avons présentée est sans surprise et assez simple, ce n'est pas le cas de la deuxième. Nous avons rencontré de nombreuses difficultés pratiques (nombre de patients disponibles limité, probabilité de faire une thrombose mal connue mais aussi lenteur des simulations...) qui nous ont retardé pour bien calibrer le test séquentiel. C'est en partie pour cela que ce deuxième test est assez mauvais et largement améliorable.

Pour finir, précisons que nous avons acquis récemment des données médicales sur les thromboses qui vont nous permettre de mettre en place le test précédent sur un cas concret et de l'optimiser en conséquence.

Références

1 [1] Robert V. Hogg, Elliot A. Tanis (2001) Probability and Statistical Inference, Prentice Hall

2 [2] Paul S. Toulouse, (1999) Thèmes de probabilités et statistique, Dunod

3 [3] Richard Durrett (1996) Probability theory and examples, Duxbury Press International Thomson Publishing

4 [4] Sheldon M. Ross (2003) Introduction to probability models, Academic Press